Tianyu Zhang · Oliver Lester Saldanha ·
Luyi Han · Nika Rasoolzadeh ·
Lidia Garrucho Moras · Jarek van Dijk ·
Tao Tan · Jakob Nikolas Kather ·
Ritse Mann (Eds.)

LNCS 16142

Artificial Intelligence and Imaging for Diagnostic and Treatment Challenges in Breast Care

**Second Deep Breast Workshop, Deep-Breath 2025
Held in Conjunction with MICCAI 2025
Daejeon, South Korea, September 23, 2025, Proceedings**



Springer

MOREMEDIA

Lecture Notes in Computer Science **16142**

Founding Editors

Gerhard Goos
Juris Hartmanis

Editorial Board Members

Elisa Bertino, *Purdue University, West Lafayette, IN, USA*
Wen Gao, *Peking University, Beijing, China*
Bernhard Steffen, *TU Dortmund University, Dortmund, Germany*
Moti Yung, *Columbia University, New York, NY, USA*

The series Lecture Notes in Computer Science (LNCS), including its subseries Lecture Notes in Artificial Intelligence (LNAI) and Lecture Notes in Bioinformatics (LNBI), has established itself as a medium for the publication of new developments in computer science and information technology research, teaching, and education.

LNCS enjoys close cooperation with the computer science R & D community, the series counts many renowned academics among its volume editors and paper authors, and collaborates with prestigious societies. Its mission is to serve this international community by providing an invaluable service, mainly focused on the publication of conference and workshop proceedings and postproceedings. LNCS commenced publication in 1973.

Tianyu Zhang · Oliver Lester Saldanha ·
Luyi Han · Nika Rasoolzadeh ·
Lidia Garrucho Moras · Jarek van Dijk · Tao Tan ·
Jakob Nikolas Kather · Ritse Mann
Editors

Artificial Intelligence and Imaging for Diagnostic and Treatment Challenges in Breast Care

Second Deep Breast Workshop, Deep-Breath 2025
Held in Conjunction with MICCAI 2025
Daejeon, South Korea, September 23, 2025
Proceedings

Springer

Editors
Tianyu Zhang
Radboudumc
Nijmegen, The Netherlands

Luyi Han
Radboudumc
Nijmegen, The Netherlands

Lidia Garrucho Moras
University of Barcelona
Barcelona, Spain

Tao Tan
Macao Polytechnic University
Macao, China

Ritse Mann
Radboudumc
Nijmegen, The Netherlands

Oliver Lester Saldanha
Dresden University of Technology
Dresden, Germany

Nika Rasoolzadeh
Radboudumc
Nijmegen, The Netherlands

Jarek van Dijk
Radboudumc
Nijmegen, The Netherlands

Jakob Nikolas Kather
Dresden University of Technology
Dresden, Sachsen, Germany

ISSN 0302-9743 ISSN 1611-3349 (electronic)
Lecture Notes in Computer Science
ISBN 978-3-032-05558-3 ISBN 978-3-032-05559-0 (eBook)
https://doi.org/10.1007/978-3-032-05559-0

This work was supported by European Federation for Cancer Images (101100633), Initiative Krebsforschung (2022 - UE77104006) and Vienna Science and Technology Fund (PREDICTOME - 10.47379/LS20065).

© The Editor(s) (if applicable) and The Author(s), under exclusive license
to Springer Nature Switzerland AG 2026
Chapter "LesiOnTime – Joint Temporal and Clinical Modeling for Small Breast Lesion Segmentation in Longitudinal DCE-MRI" is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/). For further details see license information in the chapter.

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

If disposing of this product, please recycle the paper.

Preface

The 2nd Workshop on AI and Imaging for Diagnostic and Treatment Challenges in Breast Care (Deep-Breath 2025) was held in Daejeon, South Korea, on September 23, 2025, as a full-day event at MICCAI 2025.

The Deep-Breath workshop provides an international platform for the presentation of – and discussion of – studies related to artificial intelligence in breast imaging. By fostering collaboration between clinicians and AI experts, the workshop promotes the development and responsible application of AI technologies across the breast cancer care continuum. Deep-Breath aims to create a unique forum where technical advances and clinical insights converge, driving innovation that ultimately leads to real-world impact for both physicians and patients.

Following the success of its inaugural edition in 2024, the 2025 workshop received 62 paper submissions and organized three associated international challenges – the ODELIA Challenge, MAMA-MIA Challenge, and UUSIC Challenge – which together attracted over 300 participants. Altogether, Deep-Breath 2025 engaged a truly global community, with contributions from 34 countries.

All paper submissions were double-blindly peer-reviewed by at least two domain experts. After a rigorous selection process, 35 papers were accepted for presentation at the workshop and are included in these proceedings, showcasing cutting-edge research across various imaging modalities and clinical tasks.

We would like to take this opportunity to express our sincere gratitude to MICCAI for providing an excellent platform. Special thanks to our sponsors – including ODELIA, SAFE-MRI, RadioVal, and BIG – for their generous support. We also thank our dedicated organizing team, advisory board, technical and clinical committees, as well as all authors, reviewers, and challenge participants for their invaluable contributions to the success of Deep-Breath 2025.

September 2025

Tianyu Zhang
Oliver Lester Saldanha
Luyi Han
Nika Rasoolzadeh
Lidia Garrucho Moras
Jarek van Dijk
Tao Tan
Jakob Nikolas Kather
Ritse Mann

Organization

General Chairs

Tianyu Zhang	Radboudumc, the Netherlands
Oliver Lester Saldanha	Dresden University of Technology, Germany
Luyi Han	Radboudumc, the Netherlands
Nika Rasoolzadeh	Radboudumc, the Netherlands
Lidia Garrucho Moras	Universitat de Barcelona, Spain
Jarek van Dijk	Radboudumc, the Netherlands
Tao Tan	Macao Polytechnic University, China
Jakob Nikolas Kather	Dresden University of Technology, Germany
Ritse M. Mann	Radboudumc, the Netherlands

Chairs of the Three Associated Challenges

ODELIA 2025

Nicholas Payne	University of Cambridge, UK

MAMA-MIA 2025

Lidia Garrucho Moras	Universitat de Barcelona, Spain

UUSIC 2025

Tao Tan	Macao Polytechnic University, China

Organizing Committee

Tianyu Zhang	Radboudumc, the Netherlands
Oliver Lester Saldanha	Dresden University of Technology, Germany
Nika Rasoolzadeh	Radboudumc, the Netherlands
Gustav Müller-Franzes	Aachen University Hospital, Germany
Nicholas Payne	University of Cambridge, UK
Luyang Luo	Harvard Medical School, USA

Yuan Gao	University Medical Center Utrecht, the Netherlands
Michael Kalogeropoulos	Hygeia Hospital, Greece
Robert Marti Marly	University of Girona, Spain
Daniel Truhn	University Hospital RWTH Aachen, Germany
Shandong Wu	University of Pittsburgh, USA
Oliver Diaz	University of Barcelona, Spain
Tao Tan	Macao Polytechnic University, China
Jakob Nikolas Kather	Dresden University of Technology, Germany
Ritse M. Mann	Radboudumc, the Netherlands

Technical Committee

Behdad Dasht Bozorg	Netherlands Cancer Institute, the Netherlands
Lishan Cai	Netherlands Cancer Institute, the Netherlands
Jarek van Dijk	Radboudumc, the Netherlands
Ming Fan	Hangzhou Dianzi University, China
David Friedrich	StratifAI, Germany
Luyi Han	Radboudumc, the Netherlands
Firas Khader	StratifAI, Germany
Xinglong Liang	Radboudumc, the Netherlands
Zehui Lin	Macao Polytechnic University, China
Aitor Lopez	Ribera Salud Hospital, Spain
Chunyao Lu	Radboudumc, the Netherlands
Adrià Marcos	Vall d'Hebron Institute of Oncology, Spain
Lidia Garrucho Moras	University of Barcelona, Spain
Lars Ole Schwen	Fraunhofer Institute for Digital Medicine MEVIS, Germany
Xin Wang	Netherlands Cancer Institute, the Netherlands
Markus Wenzel	Fraunhofer Institute for Digital Medicine MEVIS, Germany
Jian Zheng	Chinese Academy of Sciences, China
Jiefu Zhu	Dresden University of Technology, Germany

Clinical Committee

Alexandra Athanasiou	Mitera Hospital, Greece
Lingyun Bao	Westlake University, China
Julia Camps	Ribera Salud Hospital, Spain

Anna D'Angelo	Fondazione Policlinico Universitario A. Gemelli IRCCS, Italy
Fiona Gilbert	University of Cambridge, UK
Ritse M. Mann	Radboudumc, the Netherlands
Ning Mao	Qingdao University, China
Raquel Perez-Lopez	Vall d'Hebron Institute of Oncology, Spain
Antonio Portaluri	University of Messina, Italy
Ruixiang Qi	Westlake University, China / Netherlands Cancer Institute, Netherlands
Cristina Rossi	University of Zürich, Switzerland
Wendelien Sanderink	Radboudumc, the Netherlands
Carla Sitges	Hospital Clínic de Barcelona, Spain
Marialena Tsarouchi	Radboudumc, the Netherlands
Wouter Veldhuis	University Medical Center Utrecht, the Netherlands
Dong Xu	Chinese Academy of Sciences, China

Advisory Board

Hao Chen	Hong Kong University of Science and Technology, China
Peter Gordebeke	European Institute for Biomedical Imaging Research, Austria
Karim Lekadir	Universitat de Barcelona, Spain
Shuo Li	Case Western Reserve University, USA
Geert Litjens	Radboudumc, the Netherlands
Ulrike Mayerhofer-Schirmer	European Institute for Biomedical Imaging Research, Austria

Session Moderators

Yaofei Duan	Macao Polytechnic University, China/Netherlands Cancer Institute, the Netherlands
Gustav Müller-Franzes	Aachen University Hospital, Germany
Yuan Gao	University Medical Center Utrecht, the Netherlands
Luyi Han	Radboudumc, the Netherlands
Xinglong Liang	Radboudumc, the Netherlands
Chunyao Lu	Radboudumc, the Netherlands
Hatem A. Rashwan	Universidad Rovira i Virgili, Spain

Nika Rasoolzadeh	Radboudumc, the Netherlands
Oliver Lester Saldanha	Dresden University of Technology, Germany
Xin Wang	Netherlands Cancer Institute, the Netherlands
Tianyu Zhang	Radboudumc, the Netherlands

Additional Reviewers

Haoming Chen	Yuhan Wang
Haoran Dou	Xinghe Xie
Yaofei Duan	Xinyu Xie
Yunzhi Huang	Yi Xin
Smriti Joshi	Peirong Xu
Kaisar Kushibar	Qiuhui Yang
Xinyu Ma	Zhikai Yang
Junqiang Ma	Zhuoneng Zhang
Dandan Qiu	Zhengbo Zhou

Contents

SynBT: High-Quality Tumor Synthesis for Breast Tumor Segmentation by 3D Diffusion Model 1
Hongxu Yang, Edina Timko, Levente Lippenszky, Vanda Czipczer, and Lehel Ferenczi

Mammo-SAE : Interpreting Breast Cancer Concept Learning with Sparse Autoencoders 11
Krishna Kanth Nakka

Training Set Diversity: A Key Factor in AI-Driven Breast Ultrasound Classification 21
Rebecca Mes, Mark Wijkhuizen, Lennard M. van Karnenbeek, Tao Tan, Ritse Mann, Theo Ruers, Freija Geldof, and Behdad Dashtbozorg

FedBC: Privacy-Preserving Breast Cancer Diagnosis from Ultrasound Images Using Federated Learning 31
Faisal Ahmed, David Sánchez, Josep Domingo-Ferrer, and Zouhair Haddi

Towards Robust Breast Segmentation: Leveraging Depth Awareness and Convexity Optimization For Tackling Data Scarcity 41
Mohammad Hossein Zolfagharnasab, Tiago Gonalves, Pedro Ferreira, Maria J. Cardoso, and Jaime S. Cardoso

Divergence-Aware Training with Automatic Subgroup Mitigation for Breast Tumor Segmentation 52
Eleonora Poeta, Luisa Vargas, Daniele Falcetta, Vincenzo Marciano', Eliana Pastor, Tania Cerquitelli, Elena Baralis, and Maria A. Zuluaga

Towards Early Detection: AI-Based Five-Year Forecasting of Breast Cancer Risk Using Digital Breast Tomosynthesis Imaging 63
Manon A. Dorster, Felix J. Dorfner, Mason C. Cleveland, Melisa S. Guelen, Jay Patel, Dania Daye, Jean-Philippe Thiran, Albert E. Kim, and Christopher P. Bridge

Large Kernel MedNeXt for Breast Tumor Segmentation and Self-normalizing Network for pCR Classification in Magnetic Resonance Images 72
Toufiq Musah

Towards Affordable Tumor Segmentation and Visualization for 3D Breast MRI Using SAM2 81
Solha Kang, Eugene Kim, Joris Vankerschaver, and Utku Ozbulak

Breast Cancer Detection from Multi-view Screening Mammograms with Visual Prompt Tuning 91
Han Chen and Anne L. Martel

Are Vision Foundation Models Ready for Out-of-the-Box Medical Image Registration? 101
Hanxue Gu, Yaqian Chen, Nicholas Konz, Qihang Li, and Maciej A. Mazurowski

Towards Breast Cancer Recurrence Prediction Using Transformer-Based Learning from Global–Local Radiomics and Clinical Data 113
Adnan Khalid, Muhammad Mursil, Fabrice Meriaudeau, Ibrahima Faye, Alain Lalande, Domenec Puig, and Hatem A. Rashwan

Computer-Assisted Surgical Planning for DIEP Flap Breast Reconstruction Surgery 123
Jakub Ceranka, Diego Lamtenzan, Pieter Boonen, Ayush Kapila, Carola Brussaard, Moustapha Hamdi, and Jef Vandemeulebroucke

MM + CD Fusion: Deep Learning-based 3D Multi-Modal Fusion for Early Pathological Complete Response Prediction in Breast Cancer 134
Tewele W. Tareke, Neree Payan, Alexandre Cochet, Alain Lalande, and Fabrice Meriaudeau

Exploring Synergies Between Convolutional Neural Networks and Transformers for Breast Cancer Segmentation 145
Carlos Santiago and Jacinto C. Nascimento

A Domain-Inspired, Semi-supervised nnU-Net Is All You Need for Primary Breast Cancer Segmentation 155
Daan Schouten, Jeong Hoon Lee, Gregory Bean, Haruka Itakura, and Mirabela Rusu

Segmenting Invasive and In Situ Carcinoma in Breast WSIs with a Pretrained Histopathology Transformer 165
Lauren Jimenez-Martin, Felip Vilardell, Ana Petit, Eduard Dorca, Teresa Soler-Monsó, Irma Ramos-Oliver, Laura Pons, Pedro L. Fernández, Laia Adalid-Llansa, Daniel Mata, Lourdes Salazar-Huayna, Josep R. Casas, and Verónica Vilaplana

SiameseOrdinalCLIP: A Language-Guided Siamese Network for the Aesthetic Evaluation of Breast Cancer Locoregional Treatment 176
Fábio Teixeira, Helena Montenegro, Eduard Bonci, Maria J. Cardoso, and Jaime S. Cardoso

Anatomically and Clinically Informed Deep Generative Model for Breast Surgery Outcome Prediction 186
Joana Santos, Helena Montenegro, Eduard Bonci, Maria J. Cardoso, and Jaime S. Cardoso

BreastSegNet: Multi-label Segmentation of Breast MRI 196
Qihang Li, Jichen Yang, Yaqian Chen, Yuwen Chen, Hanxue Gu, Lars J. Grimm, and Maciej A. Mazurowski

Anatomy-Preserving Counterfactual Edits in Breast MRI via Guided Diffusion 206
Bulat Maksudov, Kathleen M. Curran, and Alessandra Mileo

Uncertainty-Aware Cross-Modal Attention for Breast Cancer Classification in Ultrasound Imaging 216
Aiman Farooq, Chandisha Das, and Deepak Mishra

Comparing Conditional Diffusion Models for Synthesizing Contrast-Enhanced Breast MRI from Pre-contrast Images 226
Sebastian Ibarra, Javier del Riego, Alessandro Catanese, Julian Cuba, Julian Cardona, Nataly Leon, Jonathan Infante, Karim Lekadir, Oliver Diaz, and Richard Osuala

Adapting Vision Language Models for Structured Clinical Description Generation in Mammography 237
André C. Castro, Dayane Rodrigues, and Arlindo R. G. Filho

Can We Teach AI to Understand Breast Tumour Behaviour? Our MAMA-MIA Challenge Journey 248
Hadeel Awwad, Joan C. Vilanova, and Robert Martí

MM-DETR: Emulating the Diagnostic Clinical Workflow in Multi-view Multi-modal Mammography Mass Detection 258
Karim Elbarbary, Adarsh Bhandary Panambur, Sheethal Bhat, Siming Bayer, and Andreas Maier

Predicting Aesthetic Outcomes of Breast Cancer Surgery: A Robust and Explainable Image Retrieval Approach 268
Pedro Ferreira, Mohammad Hossein Zolfagharnasab, Tiago Gonçalves, Eduard Bonci, Carlos Mavioso, Maria J. Cardoso, and Jaime S. Cardoso

Towards Utilizing Robust Radiance Fields for 3D Reconstruction of Breast Aesthetics 279
Gonçalo Pinto, Mohammad Hossein Zolfagharnasab, Luís F. Teixeira, Helena Cruz, Maria J. Cardoso, and Jaime S. Cardoso

Bridging the Gap: A Community Driven and AI-Enabled Approach to Early Breast Cancer Detection in Black African Women 289
Adaobi Emegoakor, Confidence Raymond, Yewande Gbadamosi, Richard Malumba, Charity Umoren, Diana Spence Betancourt, Aondona M. Iorumbur, Chinasa Kalaiwo, Abbas Rabiu Muhammad, Dennis Musinguzi, Patience Atukunda, Peter Makhoul, Rosta Asiimwe, Alfred Jatho, Amaka Nnamani, Franca Eze, Oluyemisi Toyobo, Abiodun Fatade, Udunna C. Anazodo, Farouk Dako, Michael Kawooya, and Maruf Adewole

Mammographic Image Generation Using Generative Cellular Automata 300
Lea Schwarz, Ricardo Montoya-del-Angel, Marawan Elbatel, and Robert Marti

On Tackling Domain Shift in Breast MRI Using Only Publicly-Available Data: Reproducible Breast Cancer Segmentation and pCR Prediction 310
Jessica Kächele, Dimitrios Bounias, Alexandra Ertl, and Klaus Maier-Hein

Input Simplification Impact on Robustness for Targeted Therapy Subtypes in Breast MRI Segmentation AI 320
Alba Bernal Rodriguez, Beatriz Remeserio, Justin Engelmann, Lucas Gago, and Jacinto Velasco Rebolledo

LesiOnTime - Joint Temporal and Clinical Modeling for Small Breast Lesion Segmentation in Longitudinal DCE-MRI 329
Mohammed Kamran, Maria Bernathova, Raoul Varga, Christian F. Singer, Zsuzsanna Bago-Horvath, Thomas Helbich, Georg Langs, and Philipp Seeböck

MedMask: A Self-supervised Vision Foundation Model for Breast Cancer Detection Using Mammograms 340
Tajamul Ashraf, Suhaib Salmani, Mohsen Peerzada, Ufaq Khan, Yutong Xie, and Janib Bashir

Longitudinal Mammogram Exam-Based Breast Cancer Diagnosis Models: Vulnerability to Adversarial Attacks 351
Zhengbo Zhou, Degan Hao, Dooman Arefan, Margarita Zuley, Jules Sumkin, Lars Grimm, James Joshi, and Shandong Wu

Author Index 363



SynBT: High-Quality Tumor Synthesis for Breast Tumor Segmentation by 3D Diffusion Model

Hongxu Yang[1(✉)], Edina Timko[2], Levente Lippenszky[2], Vanda Czipczer[2], and Lehel Ferenczi[2]

[1] Science and Technology Org. AI and ML, GE HealthCare, Eindhoven, The Netherlands
Hongxu.Yang@gehealthcare.com
[2] Science and Technology Org. AI and ML, GE HealthCare, Budapest, Hungary

Abstract. Synthetic tumors in medical images offer controllable characteristics that facilitate the training of machine learning models, leading to an improved segmentation performance. However, the existing methods of tumor synthesis yield suboptimal performances when tumor occupies a large spatial volume, such as breast tumor segmentation in MRI with a large field-of-view (FOV), while commonly used tumor generation methods are based on small patches. In this paper, we propose a 3D medical diffusion model, called SynBT, to generate high-quality breast tumor (BT) in contrast-enhanced MRI images. The proposed model consists of a patch-to-volume autoencoder, which is able to compress the high-resolution MRIs into compact latent space, while preserving the resolution of volumes with large FOV. Using the obtained latent space feature vector, a mask-conditioned diffusion model is used to synthesize breast tumors within selected regions of breast tissue, resulting in realistic tumor appearances. We evaluated the proposed method for a tumor segmentation task, which demonstrated the proposed high-quality tumor synthesis method can facilitate the common segmentation models with performance improvement of 2–3% Dice Score on a large public dataset, and therefore provides benefits for tumor segmentation in MRI images.

Keywords: Synthetic tumor · breast tumor segmentation · MRI

1 Introduction

Tumor synthesis provides the possibility of artificially generating the tumor examples in medical images [4]. This method is valuable when there is insufficient number of training data for deep learning model, either without considerable amount of accurate voxel-level annotation, or with low variations in tumor characteristics. To train a robust deep learning model, training samples should typically cover a large characterization space, such as imaging protocol, patient demographics, tumor characteristics and image artifacts. These challenges drastically increase the efforts for data collection, cleaning and annotation, which increase the cost of model development targeting a single tumor type [3].

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
T. Zhang et al. (Eds.): Deep-Breath 2025, LNCS 16142, pp. 1–10, 2026.
https://doi.org/10.1007/978-3-032-05559-0_1

In recent years, with the rapid achievements of image generation methods in deep learning, especially Generative Adversarial Networks (GAN) and Denoising Diffusion Probabilistic Model (DDPM) [7,12], medical images can be synthetically generated with high fidelity and valid clinical context [6,9,12]. Nevertheless, these methods require a large amount of training images to generate volume-level medical images, making both data collection and model training expensive. In contrast, Chen et al. [3] proposed a different strategy for image synthesis. The method [3] generates the tumor on a healthy image as the background, which reduces the effort of data collection and model training. However, this tumor synthesis method [3] is designed only for early stage tumors in small patches, and is not feasible for breast tumor in MRI images with a typical volume size of 300^3 voxels, therefore its generalization for different tumor types is compromised.

In this work, we introduce a novel framework, called SynthBT to generate high-quality breast tumor in contrast-enhanced MRI images, which is able to boost the tumor segmentation performance on a public MRI dataset [5]. Specifically, the proposed generation framework for breast tumor segmentation consists of three major components: 1) A two-step patch-to-volume Vector Quantized Autoencoder (VQ-VAE) [13] training. Inspired by 3D MedDiffusion [14], we implement a two-stage VQ-VAE training strategy to preserve the high-quality volumetric information of common MRI images. 2) Training a diffusion model in the latent space, that is guided by the provided tumor masks. The task of the generative model is to inpaint the removed tumor/healthy regions, with synthetic textures. 3) Training 3D segmentation models with large FOV, which is able to keep the semantic information for whole breast area, and to provide stable segmentation of the tumors. With controlling the tumor position and size, the diffusion model can produce a realistic tumor synthesis results, which leads to better segmentation performance on the considered open-sourced dataset.

The key contributions of this paper are two-fold. First, we introduce a high-quality tumor synthesis method for high-resolution breast MRI images. Extensive validation on the public dataset demonstrates that our method is effective and capable of enhancing performance in the targeted tumor segmentation tasks. Second, our work is the first to implement and validate a synthetic tumor generation framework on a large-scale public dataset with high variability in image quality for breast tumor segmentation in contrast MRI imaging, which shows the possible future directions for using synthetic data in AI model training for breast cancer applications.

2 Methods

The proposed synthetic breast tumor generation and segmentation framework consists of three different models, as shown in Fig. 1 and Fig. 2. The first stage is a two-steps VQ-VAE to efficiently compress the observed volumetric data into a latent space, while preserving the high-quality MRI volumes with large FOV. The second step is to train a latent space diffusion to synthetically generate the breast tumor based on the conditioning mask and the input volume. The final step is to train a semantic segmentation model to segment the breast tumor on MRI images by using synthetic images as supplementary data.

2.1 Patch-to-Volume VQ-VAE

The state-of-the-art tumor synthesis model [3] is designed to generate small tumor in patch-based segmentation model for CT images, which typically generates the target in a small patch of size 96^3 voxels. Nevertheless, due to high-resolution of the breast DCE-MRI images, this small patch size fails to capture sufficient semantic information, which can compromise the segmentation performance. To address this, we consider using larger volumetric inputs. However, applying diffusion models directly in the image domain for large volumes incurs high computational costs and training instability. Therefore, diffusion in the latent space is preferred for 3D medical image generation [9,14]. To obtain the latent space information, VQ-VAE is the most practical choice because of its highly efficient and low-cost design. Despite these advances, compressing a large volumetric data remains challenging for its memory consumption and training difficulties as the number of observed data points increases [14]. To overcome these limitations, 3D MedDiffusion [14] introduces a two-step region-to-global strategy for a high-quality VQ-VAE training by first learning local representations and then aggregating them into a global context.

In our implementation of Patch-to-Volume VQ-VAE, we propose to use multi-sequence contrast-enhanced MRI as the input data. Specifically, the pre-contrast MRI $\mathbf{X}_{pre}$ and a post-contrast MRI $\mathbf{X}_{post}$ are combined as the input for the compression model: $\mathbf{X} = \text{Concat}(\mathbf{X}_{pre}, \mathbf{X}_{post}) \in \mathbb{R}^{H\times W\times D\times 2}$. This combined input is divided into small patches $\mathbf{x}_i \in \mathbb{R}^{h\times w\times d\times 2}$, which are processed by the encoder $E(\cdot)$ of the VQ-VAE to produce the latent vectors $\mathbf{z}_i = E(\mathbf{x}_i)$. These latent vectors are then reassembled based on their original grid position to form the latent volume $\mathbf{Z}$, which are then processed by element-wise quantization from a spatial-based codebook $\mathcal{C} = \{\mathbf{e}_k\}_{k=1}^{K}$, where $\mathbf{K}$ is codebook size. The quantization operates each spatial code z_i to its nearest vector e_k. The decoder $D(\cdot)$ reconstructs the volume from the quantized latent volume $\hat{\mathbf{Z}}$, as $\hat{\mathbf{X}} = D(\hat{\mathbf{Z}})$.

During the volume-level tuning, only the decoder is finetuned while keeping the encoder fixed to avoid the grid artifacts between different patches and to ensure a high-quality reconstruction. With this strategy, the training efforts for the encoder and decoder can be reduced, while encoder will only focus on the regional information extraction [14].

2.2 Latent Space Diffusion Model For Synthetic Tumor

One of the commonly used approach for synthetic tumor generation in medical images is to synthetically embed the generated tumor regions onto healthy anatomical backgrounds, with carefully defined position, size, and texture. With limited clinical data under precise definition of tumor, we will not focus on the generation of background tissue, such as dense breast tissue. Instead, our method only models the foreground tumor with different varieties of its properties.

In the proposed method, with a pair of tumor-present DCE-MRI volume and tumor mask $\mathbf{M}$, the diffusion model is trained to recover the tumor texture information by given the condition of $\mathbf{Z}_{\text{cond}} = \text{cat}(\mathbf{Z}_{\text{masked}}, \text{down}(\mathbf{M}))$, where

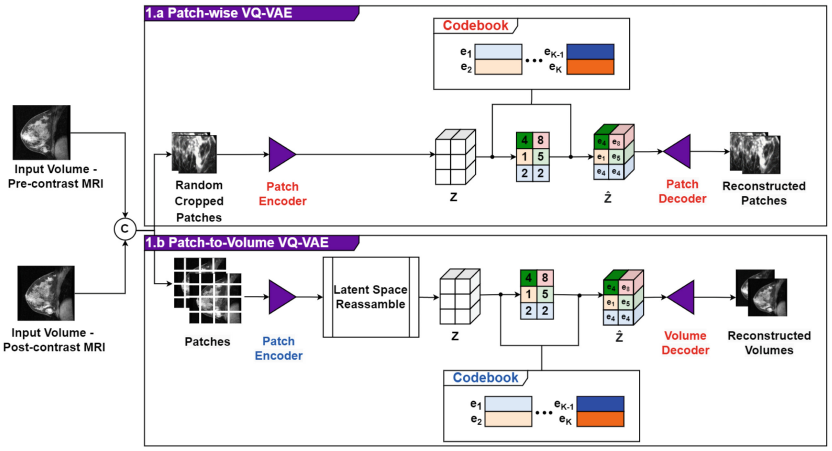


Fig. 1. Patch-to-Volume VQ-VAE training has two steps. First, the VQ-VAE is trained on random sub-patch from the input MRI images (two sequences with one pre-contrast MRI and a random post-contrast MR image). The VQ-VAE is trained to compress and reconstruct the input information in the latent space with a quantization codebook. Second, the input volumes are divided into small patches, which are processed by patch-level encoder. The compressed latent vectors are reassembled to form the latent volume based on positional order, which are then processed by the codebook, a learnable quantization indexing method, and decoder. In the first stage, the VQ-VAE is fully optimized, while the second stage will only finetune the decoder for full volume with fixed encoder and codebook. Red: trainable. Blue: frozen. (Color figure online)

down$(\cdot)$ is the down-sampling for mask to fit latent space size, while $\mathbf{Z}_{\text{masked}}$ is latent vector of the masked input volume (i.e., tumor removed). The forward diffusion process gradually adds Gaussian noise to the latent volume by T steps:

$$q(\mathbf{z}_t \mid \mathbf{z}_{t-1}) = \mathcal{N}(\mathbf{z}_t; \sqrt{1-\beta_t}\mathbf{z}_{t-1}, \beta_t \mathbf{I}) \tag{1}$$

where β_t is a variance schedule that controls the amount of Gaussian noise added at each step. The reverse process is learned via a neural network ϵ_θ, which predicts the noise added at each step, conditioned on the latent volume:

$$\mathbf{z}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{z}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(\mathbf{z}_t, t, \mathbf{Z}_{\text{cond}})\right) + \sqrt{\beta_t}\mathbf{n}, \quad \mathbf{n} \sim \mathcal{N}(0, \mathbf{I}) \tag{2}$$

where $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$ represent the cumulative product of noise retention factors over time [7]. This iterative denoising process reconstructs the tumor texture within the masked region in latent space guided by $\mathbf{Z}_{\text{cond}}$. Finally, the denoised latent volume $\hat{\mathbf{Z}}$ is passed through the VQ-VAE codebook and decoder to reconstruct the full-resolution image with the synthetic tumor.

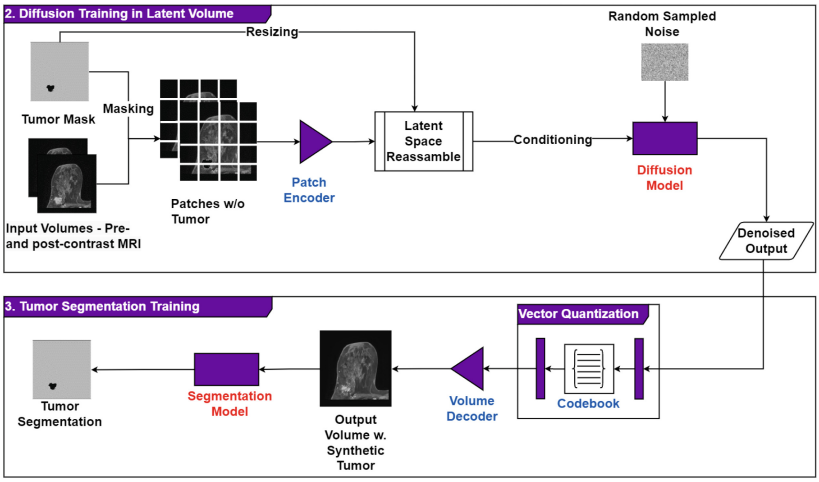


Fig. 2. With trained Patch-to-Volume VQ-VAE, a latent space diffusion model is trained for synthetic tumor generation. In this stage, tumors in the volumes are masked by ground truth, which are processed by the encoder of VQ-VAE to generate latent space vectors without tumor. Tumors are synthetically generated by diffusion model from the conditioned binary mask. The final stage of the framework is the training of the segmentation models, which is optimized by real-tumor image together with synthetically generated tumor images. Red: trainable. Blue: frozen. (Color figure online)

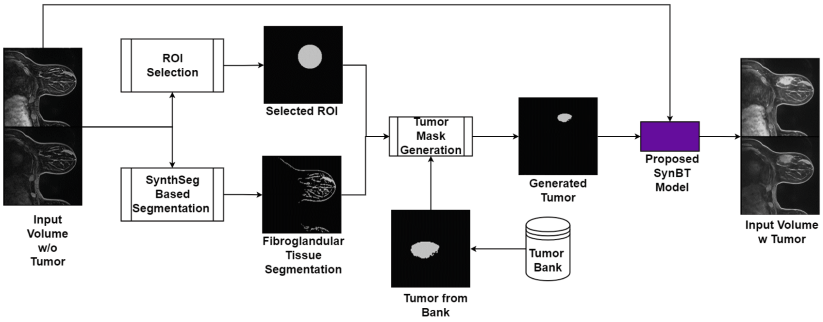


Fig. 3. The synthetic tumor generation. With given input pre- and post-contrast MRIs, the tumor mask template is selected from the tumor template bank. The tumor mask template is processed to generate final tumor mask by combining with ROI selection and fibroglandular tissue segmentation, which are controlling the size, position, shape, etc. of the tumor. Finally, the synthetic tumor is generated in pre-/post-contrast MRIs.

2.3 Segmentation Model Training with Synthetic Tumor

The diffusion model is trained to generate the tumor texture by given tumor mask and background image, which is shown in Fig. 3. In our proposal, the

tumor mask is generated by a bank of template from real-tumor in contrast-enhanced MRI, which can be augmented by flipping, rotation, zoom, etc. to generate different possible tumor masks. In contrast, the tumor is positioned to closely align with the fibroglandular tissue within the breast. Its placement and size are determined based on both the tumor and breast dimensions, ensuring that the tumor is realistically located within the breast tissue. With the generated tumor mask from the rules defined above, and the properly selected background breast MRIs, the breast tumor can be synthesized in the pre- and post-MRI images with realistic appearances. Finally, the 3D segmentation model is jointly trained with real-tumor and on-the-fly generated synthetic tumor.

3 Experimental Results

3.1 Materials and Implementation Details

Dataset: We consider the large-scale, public MAMA-MIA dataset [5] for the experiment, which consists of DCE-MRI images with carefully annotated breast tumor. To conduct the experiment, 1200 images are used for training the VQ-VAE, diffusion model and segmentation models based on the official split. The remaining 306 images are further separated to 100 images for validation and 206 images for testing. The images are isotopically resampled to 1^3 mm based on suggestion [5]. Specifically, due to multiple post-contrast sequences in the dataset, a random post-contrast MRI is selected to construct the dual-channel volume during the training, while the first post-contrast sequence is selected during the validation and testing phases. With a large amount of tumor masks in the training images, we collect these masks to construct the tumor bank for the synthetic tumor generation for segmentation tasks. In addition, the fibroglandular tissue masks from Duke dataset [10] are used to train a tissue segmentation model based on SynthSeg pipeline [1], where the rest of the body parts are clustered by k-means clustering with the number of clusters randomly selected between 2 and 8. Specifically, 3D Duke fibroglandular masks are only used in training step of the SynthSeg model. Furthermore, the segmented tissue masks are post-processed to only select the most possible regions for placing a synthetic tumor. All the input volumes are randomly cropped from the resampled DCE-MRIs with size of $128 \times 192 \times 192 \times 2$ to have sufficient context and computational trade-off.

Implementation Details: For VQ-VAE, patch size is $64^3 \times 2$ voxels, which is compressed by the VQ-VAE with compression rate of 4. The codebook has 8192 codes with dimensionality of 8. The initial convolutional channel is 32, and it's doubled as the patch resolution is reducing. The training mini-batch size is 8 with learning rate equals to 1×10^{-4}. The discriminator is started after 200,000 steps. To fine-tune the volume level reconstruction, the mini-batch size is set to 1 with learning rate of 1×10^{-5} for 100,000 steps. In both trainings, the weight of the adversarial loss and the perceptual loss are set at 0.01 and 0.001 [3,14], respectively. To train a diffusion model, a specially designed UNet [3] is used for

the generation of latent space. The denoising model is trained for 500,000 iterations with a learning rate of 1×10^{-4}. The total steps of diffusion scheduler is 1,000. During inference, the sampling is accelerated to 10 steps using the DDIM method [11]. The segmentation on full volume is processed by sliding-window strategy with size of $128 \times 192 \times 192 \times 2$ voxels ($32 \times 48 \times 48 \times 8$ in latent space), where we consider three different segmentation models for segmentation based on MONAI implementation: U-Net, nnU-Net and SwinUNET [2,3]. All segmentation models are trained to get the best validation score during the training for 300 epochs with real images, while the synthetic tumors are generated on-the-fly as supplementary dataset. In addition, data augmentations of flipping, rotation, intensity scaling and shifting are considered. The learning rate is 1×10^{-4} for the training for mini-batch=1. All experiments were conducted on a DGX A100 40GB GPU. The segmentation performance is evaluated by Dice Score (DSC), 95% Hausdorff Distance (95HD) and mean surface distance (MSD) using MedPy.

3.2 Synthetic Tumor and Segmentation Performance

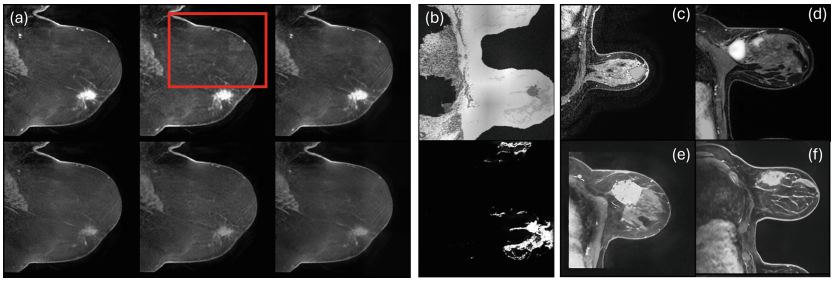


Fig. 4. (a) Left to right: real image, patch-only VQ-VAE output and Patch-to-Volume VQ-VAE outputs. Top row is post-contrast MRI, while bottom row is pre-contrast MRI. Grid effect can be observed in mid-column images for patch-only VQ-VAE. (b) SynthSeg based method for fibroglandular segmentation. Top: fake image from masks by random Gaussian Mixture Model, bottom: fibroglandular tissue as mask. (c–d) Synthetic tumor from SynTumor. (e-f) Synthetic tumor from SynBT.

The performance and ablation study of SynBT on the segmentation task are summarized in Table 1. Specifically, three different commonly considered segmentation models are reported in the table. The patch-* represents the segmentation models that are trained on small patches of 64^3 voxels, which indicates the importance of having a large field-of-view (FOV) for tumor segmentation in high-resolution DCE-MRI images. As one can observe, the patch-based segmentation model cannot handle the complex context of the breast region. In contrast, using a larger volume size can capture richer semantic information, leading to improved segmentation performance. This highlights the importance of generating high-quality synthetic data to support such methods. To compare

Table 1. Performance comparisons of segmentation results by using Dice Score (DSC), 95% Hausdorff Distance (95HD) and mean surface distance (MSD) on testing volumes.

Method	DSC (%) ↑	95HD (mm) ↓	MSD (mm) ↓
patch-UNet	0.3907	136.83	75.80
Volume-UNet	0.7220	60.05	23.90
SynTumor-UNet	0.6980	63.28	20.34
SynBT-UNet	**0.7326**	**53.79**	**17.71**
patch-nnUNet	0.3758	153.15	81.81
Volume-nnUNet	0.7463	30.77	12.48
SynTumor-nnUNet	0.7266	37.73	16.24
SynBT-nnUNet	**0.7629**	**29.26**	**10.72**
patch-SwinUNETR	0.4667	118.41	64.37
Volume-SwinUNETR	0.7278	45.25	16.65
SynTumor-SwinUNETR	0.6750	65.11	29.22
SynBT-SwinUNETR	**0.7580**	**33.99**	**11.60**

the different methods of tumor texture generation, instead of considering the diffusion-based method, we also compare the state-of-the-art analytical tumor synthesis model, SynTumor [8]. In our experiments, we found that it has a worse performance for the breast tumor synthesis, as its texture generator is not capable of generating realistic breast tumors with contrast enhanced appearance, as shown in Fig. 4. Finally, the proposed synthetic tumor generation method (SynBT-) consistently enhances the performance across various segmentation models and metrics, demonstrating its effectiveness in improving tumor segmentation.

4 Conclusions and Discussions

In this work, we propose a synthetic breast tumor generation workflow for DCE-MRI imaging, which demonstrates its potential to improve the performance of segmentation models. The method includes several key steps for segmenting breast cancer, namely Patch-to-Volume image compression, mask-guided tumor generation and tumor segmentation in DCE-MRI. The method is validated on a large-scale public dataset. The validation highlights its ability to generate realistic synthetic tumors and shows improvement for segmentation performance.

Nevertheless, there are still several limitations that could be addressed in future work. For instance, one direction could be the work for more precise placement of the tumor, due to its challenging nature in breast cancer. Our work is based on the assumption that the tumors generally exist in randomized location at the fibrograndular tissue of the breast. In clinical practice, this position information can be more precisely defined, which could lead to a potential increase in the semantic definition of the synthetic tumor. Additionally, it would

be beneficial to improve the tissue segmentation model, as it requires high efforts from post-processing side. Moreover, the diffusion based model has limitations of correctly controlling the texture with desired behavior, such as controlling the tumor categories, dedicated research could be conducted to improve the controllability of the tumor texture. Finally, dedicated quantitative synthetic image quality analysis should be conducted in the future, such as using FID, SSIM and similarity of tumor Radiomics features.

Acknowledgments. This research was supported by SYNTHIA, a project supported by the Innovative Health Initiative Joint Undertaking (IHI JU) under grant agreement No 101172872. The JU receives support from the European Union's Horizon Europe research and innovation programme. Disclaimer: Funded by the European Union, the private members, and those contributing partners of the IHI JU. The views and conclusions expressed in this paper are those of the author(s) only and do not necessarily reflect those of the aforementioned parties. Neither of the aforementioned parties can be held responsible for them.

Disclosure of Interests. This work was conducted by the authors as part of their full-time employment at GE HealthCare.

References

1. Billot, B., et al.: SynthSeg: segmentation of brain MRI scans of any contrast and resolution without retraining. Med. Image Anal. **86**, 102789 (2023)
2. Cardoso, M.J., et al.: MONAI: an open-source framework for deep learning in healthcare. arXiv preprint arXiv:2211.02701 (2022)
3. Chen, Q., et al.: Towards generalizable tumor synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11147–11158 (2024)
4. Chen, R.J., Lu, M.Y., Chen, T.Y., Williamson, D.F., Mahmood, F.: Synthetic data in machine learning for medicine and healthcare. Nat. Biomed. Eng. **5**(6), 493–497 (2021)
5. Garrucho, L., et al.: MAMA-MIA: a large-scale multi-center breast cancer dce-mri benchmark dataset with expert segmentations. arXiv e-prints, pp. arXiv–2406 (2024)
6. Guo, P., et al.: MAISI: medical AI for synthetic imaging. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 4430–4441. IEEE (2025)
7. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Adv. Neural. Inf. Process. Syst. **33**, 6840–6851 (2020)
8. Hu, Q., et al.: Label-free liver tumor segmentation. In: Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition, pp. 7422–7432 (2023)
9. Khader, F., et al.: Denoising diffusion probabilistic models for 3d medical image generation. Sci. Rep. **13**(1), 7303 (2023)
10. Saha, A., et al.: Dynamic contrast-enhanced magnetic resonance images of breast cancer patients with tumor locations [data set]. The Cancer Imaging Archive, Bethesda, MD, USA (2021)

11. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
12. Sun, L., Chen, J., Xu, Y., Gong, M., Yu, K., Batmanghelich, K.: Hierarchical amortized GAN for 3D high resolution medical image synthesis. IEEE J. Biomed. Health Inform. **26**(8), 3966–3975 (2022)
13. Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. Adv. Neural Inf. Process. Syst. **30** (2017)
14. Wang, H., Liu, Z., Sun, K., Wang, X., Shen, D., Cui, Z.: 3D MedDiffusion: a 3D medical latent diffusion model for controllable and high-quality medical image generation. IEEE Trans. Med. Imaging (2025)

Mammo-SAE: Interpreting Breast Cancer Concept Learning with Sparse Autoencoders

Krishna Kanth Nakka(✉)

Munich, Bavaria, Germany
krishkanth.92@gmail.com
https://krishnakanthnakka.github.io

Abstract. Interpretability is critical in high-stakes domains such as medical imaging, where understanding model decisions is essential for clinical adoption. In this work, we introduce Sparse Autoencoder (SAE)-based interpretability to breast imaging by analyzing Mammo-CLIP, a vision–language foundation model pretrained on large-scale mammogram image–radiology report pairs. We train a patch-level `Mammo-SAE` on Mammo-CLIP visual features to identify and probe latent neurons associated with clinically relevant breast concepts such as *mass* and *suspicious calcification*. We show that top-activated class-level latent neurons often tend to align with ground-truth regions, and also uncover several confounding factors influencing the model's decision-making process. Furthermore, we demonstrate that finetuning Mammo-CLIP leads to larger separation in the latent neuron space, improving interpretability and predictive performance. Our findings suggest that sparse latent representations offer a powerful lens into the internal behavior of breast foundation models.

Keywords: Sparse Autoencoders · Explainable AI · Breast Cancer

1 Introduction

In high-stakes domains such as healthcare, machine learning models must not only be accurate but also interpretable. To enhance transparency in breast imaging, prior work has proposed both post-hoc interpretability methods—such as GradCAM variants [8,12,18]—and inherently interpretable architectures, including those based on prototype networks [2,17]. While these tools provide useful explanations at the prediction level, they offer limited insight into the model's internal mechanisms, particularly at the level of individual neurons. Moreover, prior studies have shown that neurons in deep networks are often *polysemantic* [4,14,16], i.e., they activate in response to multiple unrelated concepts, making them difficult to interpret reliably.

Recently, Sparse Autoencoders (SAEs) [3,5,7,10,13] have gained significant traction for interpreting Large Language Models (LLMs) [20,21]. Building on

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-032-05559-0_2.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
T. Zhang et al. (Eds.): Deep-Breath 2025, LNCS 16142, pp. 11–20, 2026.
https://doi.org/10.1007/978-3-032-05559-0_2

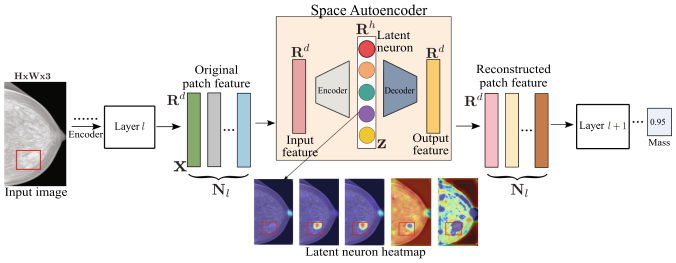


Fig. 1. Mammo-SAE Framework. The SAE is first trained on patch-level CLIP features $\mathbf{x}_j \in \mathbf{R}^d$ at any given layer, projecting them into a high-dimensional, interpretable sparse latent space $\mathbf{z} \in \mathbf{R}^h$, and decoding them back for reconstruction. Once trained, the SAE is used to analyze which latent neurons are activated and what semantic information they encode. We also perform targeted interventions in the latent neuron space to assess their influence on downstream label prediction. We observe the learned latents capture diverse regions such as *nipple regions, mass regions*, and background areas. Red box indicate ground-truth mass localization.

this progress, SAEs have also been adapted to Vision Language Models [11,19]. SAEs are capable of extracting *monosemantic* features—latent dimensions that correspond to interpretable concepts—and support test-time interventions that allow controlled probing and manipulation of model behavior. Notably, SAEs can be integrated at any layer of a model, providing a flexible and modular approach to interpreting intermediate representations. This layer-wise adaptability makes them a powerful tool for dissecting model behavior at inference time.

In this work, we extend SAE-based interpretability techniques to breast imaging by introducing `Mammo-SAE`, a sparse autoencoder trained on visual features from Mammo-CLIP [6], a vision–language foundation model pretrained on mammogram image–report pairs. Concurrent to this work, the use of SAEs in the medical domain for radiology model has also been explored in [1]. Our contributions are as follows:

1. We train a patch-based SAE on Mammo-CLIP to discover latent neurons associated with breast cancer-related concepts such as *mass* and *suspicious calcification*.
2. We uncover monosemantic features in the SAE latent space and visualize their spatial activation patterns, showing alignment with clinical regions of interest.
3. We conduct targeted group interventions on the SAE latent space to reveal that the model sometimes relies on confounding features when making decisions.
4. We find that finetuning leads to a clearer separation of latent neurons associated with breast cancer-related concepts, providing insight into observed performance gains.

2 Proposed Method

Figure 1 illustrates our proposed Mammo-SAE framework for interpreting breast foundation models; in this work, we apply it to Mammo-CLIP as a representative example. The framework consists of three main components: (i) an encoder-decoder SAE is pretrained to project CLIP features into a high-dimensional sparse latent space, encouraging disentangled and interpretable representations (Sect. 2.1); (ii) a probing framework is employed to identify latent neurons that are selectively activated in the presence of breast cancer-related concepts such as *mass* and *suspicious calcification*, enabling concept-level interpretability (Sect. 2.2); and (iii) an intervention framework is used to manipulate group of class-level latent neurons and observe changes in the model's output, allowing us to assess the causal impact of latent neurons and identify potential confounding factors influencing decision-making (Sect. 2.3).

Preliminaries. Mammo-CLIP [6] is a vision–language foundation model trained to align image and text representations using paired mammogram images and radiology reports. After pretraining, the Mammo-CLIP image embeddings can be used for downstream concept prediction (e.g., binary classification of the presence of a *mass* concept) via a single fully connected classification layer. Additionally, the model can be further finetuned by updating the CLIP backbone and classifier jointly to improve performance on specific breast concept recognition tasks. Throughout this paper, we refer to the original frozen backbone as the *pretrained* variant and the end-to-end updated model as the *finetuned* variant. We provide further details in the Supplementary material.

2.1 Mammo-SAE

Let an input image I and the feature extracted by the Mammo-CLIP model f at layer l is $\mathbf{x}$. SAE takes the Mammo-CLIP local feature $\mathbf{x}_l^j \in \mathbf{R}^d$ at layer l and spatial position j, where j indexes the spatial location in the feature map of size $N_l = H_l \times W_l$. The SAE consists of two layers: an encoder that projects the input into a high-dimensional sparse latent space, and a decoder that reconstructs the original CLIP feature from this latent representation.

The model is trained using a combination of reconstruction loss and a sparsity constraint, which encourages activation of only a small subset of neurons in the latent space, thereby enhancing interpretability. Let $W_{\text{enc}} \in \mathbf{R}^{d \times h}$ and $W_{\text{dec}} \in \mathbf{R}^{h \times d}$ denote the encoder and decoder weight matrices, respectively, and let $\phi(\cdot)$ denote the ReLU non-linear function. The training objective is defined as:

$$\mathcal{L} = \underbrace{\left\| W_{\text{dec}}\, \phi\left(W_{\text{enc}} \mathbf{x}^j\right) - \mathbf{x}^j \right\|_2^2}_{\text{ReconstructionLoss}} + \lambda \underbrace{\left\| \phi\left(W_{\text{enc}} \mathbf{x}^j\right) \right\|_1}_{\text{SparsityLoss}}, \tag{1}$$

where the first term represents the reconstruction loss, and the second term is the sparsity loss with regularization coefficient λ. Further implementation details about the SAE training can be found in Sect. 3.

2.2 Probing Mammo-SAE Latents to Identify Breast Concepts

After training Mammo-SAE on layer-l activations from Mammo-CLIP, we analyze the resulting latent space to identify neurons that correspond to specific breast cancer-related concepts in the model. Our goal is to pinpoint latent neurons that are consistently activated in the presence of concepts such as *mass* or *suspicious calcification*.

For each class label $c \in \{0, 1\}$, where $c = 1$ denotes the presence of the breast concept, we compute the class-wise mean latent activation vector $\bar{\mathbf{z}}^{(c)} \in \mathbf{R}^h$ by averaging over all spatial locations and all training samples in that class:

$$\bar{\mathbf{z}}^{(c)} = \frac{1}{|\mathcal{D}_c| \cdot N_l} \sum_{\mathbf{x} \in \mathcal{D}_c} \sum_{j=1}^{N_l} \phi\left(W_{\text{enc}}\, \mathbf{x}^j\right), \tag{2}$$

where $\mathcal{D}_c$ is the set of training images with class label c, and $\mathbf{x}^j \in \mathbf{R}^d$ denotes the CLIP feature at spatial location j in image feature $\mathbf{x}$. The function $\phi(\cdot)$ denotes the ReLU activation applied after encoding.

We then assign each latent neuron t a class-level relevance score defined as its class-specific mean activation, $s_t^{(c)} = \bar{z}_t^{(c)}$, where $\bar{z}_t^{(c)}$ is the t-th element of $\bar{\mathbf{z}}^{(c)}$. Latent neurons are ranked in descending order of $s_t^{(c)}$, and the top-scoring ones for class $c = 1$ are considered most aligned with the target concept. While we adopt a simple mean-based scoring approach, alternative strategies based on entropy or standard deviation can also be explored [11]. To assess the reliability of the top class-level latent neurons, we examine the image regions that most strongly activate each latent neuron. This analysis provides visual evidence of whether a neuron attends to clinically meaningful regions or to spurious patterns, offering insight into the model's internal reasoning.

2.3 Intervention on Mammo-SAE Latent Neurons

Furthermore, to assess the causal role of Mammo-SAE latent neurons in downstream predictions, we perform targeted interventions on the top-k class-level neurons identified for a given concept. Specifically, we introduce two types of group interventions on the patch-level SAE latent $\mathbf{z} = \phi\left(W_{\text{enc}}\, \mathbf{x}^j\right)$ at every spatial position j:

(i) **Top-k Activated:** We retain only the activations of the top-k latent neurons and zero out all others. This isolates the influence of the most concept-relevant neurons.

$$\mathbf{z}' = \mathbf{z} \odot \mathbf{m}, \quad \text{where } m_i = \begin{cases} 1, & \text{if } i \in \mathcal{T}_k^{(0)} \cup \mathcal{T}_k^{(1)} \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

Here, $\mathcal{T}_k^{(c)} \subset \{1, \ldots, k\}$ denotes the set of indices corresponding to the top-k neurons for class $c \in \{0, 1\}$, ranked by their class-specific scores $s_k^{(c)}$. A neuron position i is retained (i.e., $m_i = 1$) if it appears in the union of top-k neurons for class 0 or class 1.

Table 1. AUC-ROC comparison between the original Mammo-CLIP model and the SAE-reconstructed model. We insert SAE at the final layer.

	W/o SAE	W/ SAE
Pretrained	0.951	0.933
Finetuned	0.978	0.979

(a) Suspicious calcification

	W/o SAE	W/ SAE
Pretrained	0.786	0.763
Finetuned	0.856	0.855

(b) Mass

(ii) **Top-k Deactivated:** We zero out the top-k neurons while leaving all other latent activations unchanged. This tests the dependency of the model's prediction on these specific neurons.

$$\mathbf{z}' = \mathbf{z} \odot (1 - \mathbf{m}) \tag{4}$$

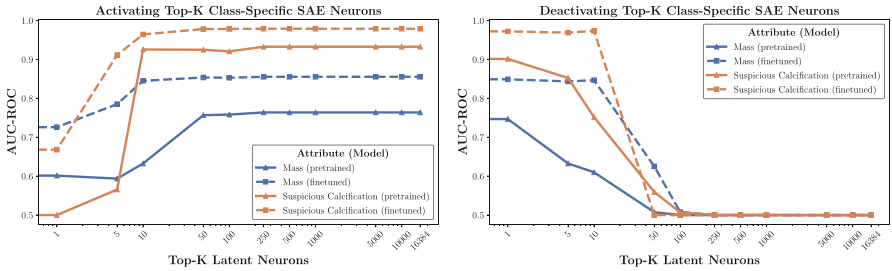
By measuring the change in the model's output before and after these interventions, we assess the functional importance of the selected neurons and determine whether the model relies on meaningful features or potentially confounding patterns.

3 Experiments

Dataset. We conduct our experiments on the VinDr-Mammo dataset [15], which contains approximately 20,000 mammogram images from 5,000 patients.

SAE Training. We utilize the Vision-SAEs library [19] to train a Sparse Autoencoder (SAE) on patch-level features extracted from the fine-tuned Mammo-CLIP model [6]. We focus specifically on the classifier trained for the *suspicious calcification* concept, using *activations from the final layer* of the EfficientNet-B5 backbone [9] of Mammo-CLIP [6]. Rather than training separate SAEs for each model which is expensive, we train a single SAE once and reuse it across all experiments. This shared SAE design not only reduces overhead but also ensures a consistent latent space, making it easier to compare representations across different models (see Sect. 3.3). The input feature dimension is $d = 2048$, and we set the expansion factor to 8, resulting in a latent dimension of $h = 16{,}384$. The SAE is trained for 200 epochs with a learning rate of 3×10^{-4}, sparsity penalty $\lambda = 3 \times 10^{-5}$, and batch size of 4096.

SAE Generalizability. In Table 1, we compare the predictive performance of models using original CLIP features versus SAE-reconstructed features, for both *mass* and *suspicious calcification* concepts. We conduct this comparison on both the pretrained and fine-tuned variants of Mammo-CLIP to assess whether the SAE preserves original information. Across both models and both concepts, we observe that the drop in AUC-ROC is less than 2%, indicating that the SAE—



(a) Top-k Activated (b) Top-k Deactivated

Fig. 2. Intervention on class-level latent neurons. Left: Top-k activated intervention—only the top-k class-specific neurons are retained, and all others are zeroed out. Right: Top-k deactivated intervention—the top-k neurons are zeroed out while the rest remain unchanged.

trained once on a single network—generalizes well and retains reliable representations for downstream prediction. We will now explore SAEs to dissect the model behaviour.

3.1 Intervention on Class-Level Latent Neurons

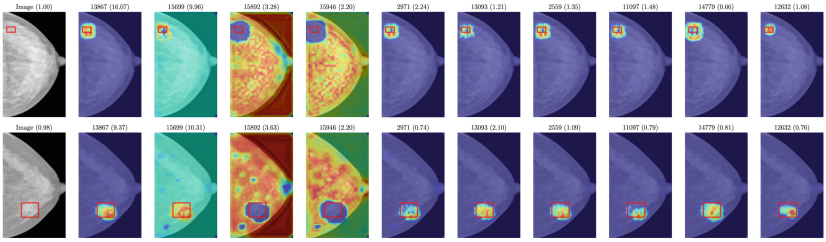
As described in Sect. 2.2, we compute the relevance score of each latent neuron with respect to two classes (e.g., *Mass* vs. *Non-Mass*) and identify the top-k neurons per class. We then perform targeted interventions as outlined in Sect. 2.3.

In the left panel of Fig. 2a, we show results for the **top-k activated** intervention, where only the top-k class-specific neurons are retained and all others are zeroed out, with k varied from 0 up to the full latent dimensionality $h = 16{,}384$. Remarkably, activating as few as 10 neurons is sufficient to nearly recover the model's original AUC-ROC in multiple cases—demonstrating that **a small subset of neurons captures most of the task-relevant signal.**

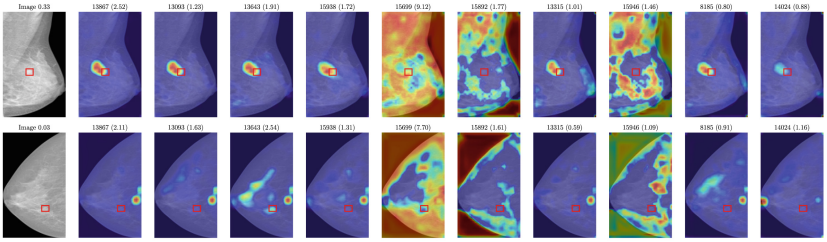
Conversely, in Fig. 2b, we present results for the **top-k deactivated** intervention, where the top-k class-specific neurons are zeroed out while the rest of the latent representation is left unchanged. We observe that deactivating more than 10 neurons leads to a sharp drop in AUC-ROC, highlighting the model's strong reliance on a compact, concept-aligned subset of latent features. These findings underscore the precision and interpretability of the Mammo-SAE representation.

3.2 Analyzing Top Activated Latent Neurons

To interpret the internal representations learned by Mammo-SAE, we visualize the activations of the top-k latent neurons from the encoded representation $\mathbf{z}$ at each spatial location j.



(a) Suspicious Calcification (Finetuned)



(b) Mass (Finetuned)

Fig. 3. Visualization of the top-10 class-level latent neurons of class $c = 1$ from the fine-tuned Mammo-SAE model for two breast cancer concepts. Red boxes denote ground-truth concept regions. Each image is annotated with the latent neuron index and its mean activation value. Best viewed in zoom. Additional examples are provided in the supplementary material.

Figures 3a and 3b show heatmaps of the top-10 latent neurons most associated with the positive class ($c = 1$) for two breast cancer concepts: *suspicious calcification* and *mass*. Ground-truth concept regions are overlaid in red for clarity. For suspicious calcification, we observe that 7 out of the top 10 latent neurons activate strongly within the annotated region, indicating that Mammo-SAE has learned semantically meaningful and spatially aligned representations. In contrast, for the mass concept, the top-activated neurons show weak alignment with ground-truth regions, which aligns with the relatively lower AUC-ROC observed in Table 1.

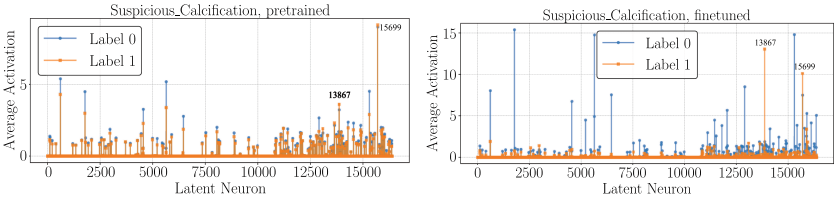
To quantitatively assess the spatial alignment between SAE latent activations and annotated breast concept regions, we threshold each latent heatmap at the 95th percentile and extract rectangular bounding boxes to approximate predicted concept locations. Table 2 reports the mean Average Precision (mAP) at an Intersection-over-Union (IoU) threshold of 0.25, computed over the top-10 class-selective latent neurons for both the pretrained and finetuned variants of Mammo-CLIP. We find that the fine-tuned model consistently achieves higher mAP than the pretrained variant, suggesting that fine-tuning enhances the model's ability to align concept-relevant features with spatially meaningful regions. Conversely, the lower mAP in the pretrained model indicates that

Table 2. Mean Average Precision (mAP) for breast concept localization using top-10 class-level latent neuron activations of class $c = 1$ across different settings.

Concept	Model	1	2	3	4	5	6	7	8	9	10
Suspicious Calcification	Finetuned	0.256	0.007	0.005	0.007	0.084	0.102	0.084	0.083	0.166	**0.278**
Suspicious Calcification	Pretrained	0.057	0.053	0.027	**0.085**	0.008	0.002	0.002	0.011	0.033	0.053
Mass	Finetuned	0.295	**0.316**	0.308	0.286	0.0	0.0	0.159	0.0	0.182	0.135
Mass	Pretrained	**0.045**	0.019	0.029	0.053	0.022	0.020	0.024	0.022	0.018	0.025

the model often relies on spurious or task-irrelevant background regions when making predictions. It is important to note that no annotated localization information is used during training for either the pretrained or finetuned models.

These findings highlight two key insights: (1) a significant fraction of latent neurons capture clinically meaningful visual concepts, and (2) some neurons still respond to irrelevant background areas yet influence the final decision. Understanding and controlling for these background-sensitive neurons could be crucial for building more robust and interpretable breast cancer detection models. Our framework provides a concrete path forward for future efforts to mitigate reliance on confounding features during both training and inference.



(a) Pretrained (b) Finetuned

Fig. 4. Mean latent activation vectors $\bar{\mathbf{z}}^{(c)}$ for each class ($c = 1$ indicates the presence of the concept) in the pretrained model (left) and finetuned model (right) for the *suspicious calcification* concept.

3.3 Latent Neuron Separation: Finetuned Vs. Pretrained

Table 1 demonstrates that finetuned models significantly outperform pretrained models in terms of AUC-ROC. To understand the underlying cause of this improvement, we analyze the class-wise mean latent vectors $\bar{\mathbf{z}}^{(c)}$ extracted from both models for *suspicious calcification* concept, visualized in Figs. 4a and 4b.

We draw three key insights from this comparison: (1) the separation between class-wise mean activations becomes significantly more pronounced in the finetuned model, suggesting that finetuning sharpens the latent space to better distinguish the presence of breast concepts; (2) Neuron 13867 emerges as a dominant signal for the *suspicious calcification* class after finetuning which is the top-1 activated latent neuron shown in Fig. 3a (second column), highlighting

that the model learns to amplify clinically meaningful features; however (3) Neuron 15699 remains persistently active across both classes in *both* models and classes, corresponding to a spurious background region (Fig. 3a, third column), which show the evidence that even with finetuning, the model partially relies on non-discriminative or confounding features.

4 Conclusion

In this paper, we introduced Mammo-SAE, a framework for uncovering breast concept representations in the Mammo-CLIP [6]. By probing the latent space, we identified neurons that are selectively activated in the presence of clinically relevant breast concepts such as *mass* and *suspicious calcification*. Through visualization and by concept localization, we observed that while some latent neurons align well with ground-truth regions, others respond to background areas—highlighting both the strengths and limitations of the learned representations. We believe Mammo-SAE provides a valuable tool for understanding the causal mechanisms within foundation models for medical imaging and opens new avenues for inference-time interventions to improve interpretability and performance in breast cancer detection.

Acknowledgments. Author sincerely acknowledges the CHUV breast tumour board, including Dr. Khalil Zaman, Dr. Assia Ifticene Treboux, Dr. Wendy Jeanneret Sozzi, Prof. Patrice Mathevet, and Dr. Avdulla Krasniqi, for their compassionate care of a close family member and for the inspiration that profoundly shaped this work.

Disclosure of Interests. The author declares no competing interests relevant to the content of this article.

References

1. Bouzid, K., Bannur, S., de Castro, D.C., Schwaighofer, A., Alvarez-Valle, J., Hyland, S.L.: Insights into a radiology-specialised multimodal large language model with sparse autoencoders. arXiv preprint arXiv:2507.12950 (2025)
2. Choukali, M.A., Amirani, M.C., Valizadeh, M., Abbasi, A., Komeili, M.: Pseudo-class part prototype networks for interpretable breast cancer classification. Sci. Rep. **14**(1), 10341 (2024)
3. Cunningham, H., Ewart, A., Riggs, L., Huben, R., Sharkey, L.: Sparse autoencoders find highly interpretable features in language models. arXiv preprint arXiv:2309.08600 (2023)
4. Dreyer, M., Purelku, E., Vielhaben, J., Samek, W., Lapuschkin, S.: PURE: turning polysemantic neurons into pure features by identifying relevant circuits. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8212–8217 (2024)
5. Gao, L., et al.: Scaling and evaluating sparse autoencoders. arXiv preprint arXiv:2406.04093 (2024)

6. Ghosh, S., Poynton, C.B., Visweswaran, S., Batmanghelich, K.: Mammo-CLIP: a vision language foundation model to enhance data efficiency and robustness in mammography. In: Linguraru, M.G., et al. (eds.) MICCAI 2024. LNCS, vol. 15012, pp. 632–642. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-72390-2_59
7. He, Z., et al.: LLaMA scope: extracting millions of features from LLaMA-3.1-8B with sparse autoencoders. arXiv preprint arXiv:2410.20526 (2024)
8. Kajala, A., Jaiswal, S., Kumar, R.: Breaking the black box: heatmapdriven transparency to breast cancer detection with EfficientNet and grad CAM. Educ. Admin. Theory Pract. **30**(5), 4999–5009 (2024)
9. Koonce, B.: EfficientNet. In: Koonce, B. (ed.) Convolutional Neural Networks with Swift for Tensorflow, pp. 109–123. Apress, Berkeley, CA (2021). https://doi.org/10.1007/978-1-4842-6168-2_10
10. Lieberum, T., et al.: Gemma scope: open sparse autoencoders everywhere all at once on Gemma 2. arXiv preprint arXiv:2408.05147 (2024)
11. Lim, H., Choi, J., Choo, J., Schneider, S.: Sparse autoencoders reveal selective remapping of visual concepts during adaptation. arXiv preprint arXiv:2412.05276 (2024)
12. Liu, S., Himel, G.M.S., Wang, J.: Breast cancer classification with enhanced interpretability: DalaresNet50 and DT grad-CAM. IEEE Access (2024)
13. Makelov, A., Lange, G., Nanda, N.: Towards principled evaluations of sparse autoencoders for interpretability and control. arXiv preprint arXiv:2405.08366 (2024)
14. Marshall, S.C., Kirchner, J.H.: Understanding polysemanticity in neural networks through coding theory. arXiv preprint arXiv:2401.17975 (2024)
15. Nguyen, H.T., et al.: VinDr-Mammo: a large-scale benchmark dataset for computer-aided diagnosis in full-field digital mammography. Sci. Data **10**(1), 277 (2023)
16. O'Mahony, L., Andrearczyk, V., Müller, H., Graziani, M.: Disentangling neuron representations with concept vectors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3770–3775 (2023)
17. Pathak, S., Schlötterer, J., Veltman, J., Geerdink, J., Keulen, M., Seifert, C.: Prototype-based interpretable breast cancer prediction models: analysis and challenges. In: Longo, L., Lapuschkin, S., Seifert, C. (eds.) xAI 2024. CCIS, vol. 2153, pp. 21–42. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-63787-2_2
18. Raghavan, K., Sivaselvan, B., Kamakoti, V.: Attention guided grad-CAM: an improved explainable artificial intelligence model for infrared breast cancer detection. Multimed. Tools Appl. **83**(19), 57551–57578 (2024)
19. Rao, S., Mahajan, S., Böhle, M., Schiele, B.: Discover-then-name: task-agnostic concept bottlenecks via automated concept discovery. In: Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G. (eds.) ECCV 2024. LNCS, vol. 15135, pp. 444–461. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-72980-5_26
20. Team, G., et al.: Gemma: open models based on Gemini research and technology. arXiv preprint arXiv:2403.08295 (2024)
21. Touvron, H., et al.: LLaMA: open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)

Training Set Diversity: A Key Factor in AI-Driven Breast Ultrasound Classification

Rebecca Mes[1,2], Mark Wijkhuizen[1], Lennard M. van Karnenbeek[1], Tao Tan[3], Ritse Mann[4,5], Theo Ruers[1,5], Freija Geldof[1], and Behdad Dashtbozorg[1,5(✉)]

[1] Image-Guided Surgery, Department of Surgery, Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX Amsterdam, The Netherlands
{m.wijkhuizen,l.karnenbeek,t.ruers,f.geldof,b.dasht.bozorg}@nki.nl
[2] Faculty of Mechanical, Maritime, and Materials Engineering, Delft University of Technology, Mekelweg 2, 2628 CD Delft, The Netherlands
[3] Faculty of Applied Sciences, Macao Polytechnic University, Rua de Luis Gonzaga Gomes, Macao, China
[4] Department of Radiology, Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX Amsterdam, The Netherlands
r.mann@nki.nl
[5] Center for Early Cancer Detection, Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX Amsterdam, The Netherlands

Abstract. Breast ultrasound (BUS) offers a low-cost, radiation-free imaging alternative for breast cancer diagnostics, particularly suitable for point-of-care use. Despite promising results from deep learning (DL) models for BUS lesion classification, most models show significant performance drops on external datasets, suggesting overfitting to dataset-specific features. This lack of generalizability is concerning, especially given disparities in breast cancer outcomes across demographic groups and the diversity of ultrasound acquisition conditions. In this study, we investigate the impact of training dataset diversity on the robustness of DL models for BUS lesion classification. We compare three model architectures: ResNet50 (CNN), MViTv2 (Vision Transformer), and MambaOut (Vision Mamba), using 8403 B-mode BUS images from ten publicly available datasets originating from seven countries. Models were evaluated under three scenarios: single-dataset training, leave-one-dataset-out (LODO), and limited-data all-source training. Our results show that performance strongly depends on training set composition, with certain datasets consistently yielding better model performance, and substantial variability in cross-dataset generalization. This study provides new insights into the design of fair and generalizable DL systems for breast cancer diagnostics.

Keywords: Breast cancer · Lesion classification · Ultrasound · Deep learning · Data diversity · Domain generalization

R. Mes and M. Wijkhuizen—These authors contributed equally to this work.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
T. Zhang et al. (Eds.): Deep-Breath 2025, LNCS 16142, pp. 21–30, 2026.
https://doi.org/10.1007/978-3-032-05559-0_3

1 Introduction

Breast cancer is the most commonly diagnosed cancer worldwide, with 2.3 million new cases annually [1]. Early detection significantly reduces mortality rates by 43% [2], and imaging plays a key role in diagnosis. Mammography is the current gold standard for breast cancer screening but suffers from limitations such as reduced sensitivity in women with dense breast tissue, discomfort for the patient, radiation exposure, and high false-positive rates leading to a high rate of unnecessary biopsies (65–85%) [3]. Ultrasound (US) imaging offers a low-cost and radiation-free alternative, particularly suitable for point-of-care (POC) settings. Recent studies have demonstrated the efficacy of ultrasound as a stand-alone breast imaging modality in the assessment of focal breast complaints, with the ability of reducing unnecessary biopsies by 40% [3]. However, its diagnostic accuracy is highly operator-dependent and can be hindered by artifacts, low contrast, and noise.

Deep learning (DL) models have shown high accuracy in automatic BUS lesion classification. The performance reaches or even outperforms radiologists [4]. Most approaches use CNNs (particularly ResNet, VGG, and DenseNet variants), with growing interest in transformer-based models and more recently, state-space models such as Mamba. Reported accuracies for CNNs range from 77% to 98%, with AUCs up to 0.996 on public datasets [5–10]. However, multiple studies highlight performance drops on private or external datasets, sometimes ¿20% in accuracy or AUC [5,7]. Ensemble and hybrid models improve robustness slightly [6,11], but dataset-specific overfitting remains a concern. Vision transformers (ViTs) such as DeiT and HoVer-Trans achieve comparable performance (accuracies of 87 to 95%, AUC up to 0.95) while offering advantages in capturing global context compared to the local receptive fields of convolutional operations in CNNs. [4,12,13]. However, ViTs often require more training data than CNNs, raising concerns about their suitability for clinical domains with limited, heterogeneous data. Recently, the Mamba architecture has emerged for general vision tasks, but its potential for medical US imaging remains largely unexplored. To date, two studies used Mamba-based models (VMamba and MedMamba) for BUS classification [14,15], showing that numerous alternative models matched or outperformed the Mamba variant. It is noteworthy that only a minority of studies explicitly evaluate models across multiple datasets, and even fewer report performance on combined or harmonized multi-center datasets (e.g., Gu et al. used data from 32 hospitals [8]).

The role of dataset diversity, both in terms of patient demographics and acquisition heterogeneity (e.g., scanner type, probe frequency, operator skill), remains largely underexplored in current research of BUS classification. Where evaluated, studies show performance discrepancies between public and private clinical datasets, suggesting overfitting to dataset-specific characteristics. This is especially concerning given known disparities in breast cancer outcomes across ethnic groups. Although black women have a 5% lower incidence rate than white women, their mortality rate is 38% higher [16], and women with dense breast tissue have seen less reduction in mortality rates over recent years [17]. Patient-specific factors and acquisition-related biases can systematically influence AI

Table 1. Characteristics of public BUS datasets. A dash (–) denotes not reported.

	Name	US machine	Image dimension	Freq.	N_{total}	N_{benign}	$N_{malignant}$	Country	Ground truth
D1	GDPH & SYSUCC [12]	Hitachi Ascendus, Mindray DC-80, Toshiba Aplio 500, Supersonic Aixplorer	844 × 624	–	2405	886	1519	China	Pathologist
D2	QAMEBI [21]	AixPlorer Ultimate, Supersonic Imagine	580 × 504	5–18 MHz	232	109	123	Iran	Pathologist
D3	BUSI [22]	LOGIQ E9 & E9 Agile	600 × 480	1–5 MHz	647	437	210	Egypt	Radiologist
D4	BUS_UC [23]	Hitachi Medical	256 × 256	–	811	358	453	Pakistan	Radiologist
D5	BLDiUV [24]	LOGIQ E9, PHILIPS TIS L9-3	621 × 621	3–7 MHz	1194	456	738	China	Pathologist
D6	BUS-BRA [25]	GE Logiq 5 & 7, Toshiba Aplio 300, GE U-Systems	320 × 390	10–14 MHz	1875	1268	607	Brazil	Pathologist
D7	BUS-UCLM [26]	Siemens ACUSON S2000TM	856 × 606	–	255	169	86	Spain	Pathologist
D8	MID [27]	Philips iU22	770 × 640	–	244	122	122	Thailand	Radiologist
D9	UDIAT [28]	Siemens ACUSON Sequoia C512	530 × 450	8.5 MHz	163	109	54	Spain	–
D10	HMSS [29]	Hitachi Medical	480 × 380	–	577	261	316	the Netherlands	–

performance [18]. Without addressing these potential biases in training data, AI models risk perpetuating or even amplifying healthcare inequalities when deployed in real-world settings.

In this study, we evaluate the impact of training set diversity on the performance of DL models for BUS lesion classification. We compare three distinct architectures - ResNet50 [19], MViTv2 [20], and MambaOut [14] across both individual and diverse BUS datasets. Our aim is to assess how dataset composition influences model accuracy and generalizability across demographically and technically varied imaging data. To our knowledge, this is the first work to explicitly evaluate model robustness in light of dataset diversity within BUS classification, to provide insights for the development of fair and generalizable AI tools in future breast cancer diagnostics.

2 Methods

2.1 Datasets

B-mode breast ultrasound (BUS) images were retrieved from ten publicly available databases, denoted as D1 to D10, see Table 1 [12,21–29]. The datasets consist of 8403 images in total, categorized as benign (n = 4175) and malignant (n = 4228) lesions based on histopathological or radiological assessment. The data originates from seven countries across four continents, and encompass a variety of ultrasound systems and probes, enhancing domain diversity. Benign example images for every dataset are presented in Fig. 1. Each input image was center-cropped to obtain a square aspect ratio and subsequently resized to 224 × 224 pixels. Images were normalized to the range [0, 1]. From each dataset, 50 benign and 50 malignant cases were randomly selected as test set (except D9, limited to 25 benign/malignant). Images from the same patient were assigned exclusively to either the training or test split.

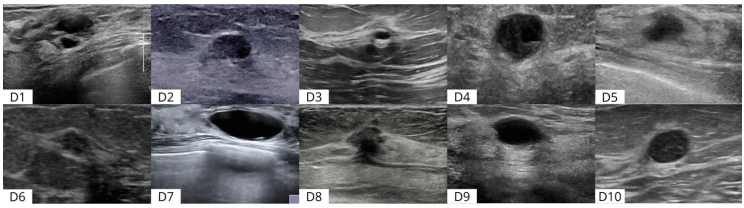


Fig. 1. Examples of different benign ultrasound images for the ten datasets.

2.2 Deep Learning Models

We selected three distinct architectures representative of different design paradigms: ResNet50 (CNN) [19], MViTv2-Small (Vision Transformer) [20], and MambaOut (Vision Mamba) [14]. All models were accessed via the PyTorch Image Models (TIMM) library, pre-trained on ImageNet1K, and with an input size of 224 × 224 pixels. Model selection aimed to maintain parameter equality (25M–100M) to ensure fair comparison.

2.3 Model Training and Data Augmentation

The models were initialized with pre-trained ImageNet1K weights and fine-tuned on the BUS datasets. No model layers were frozen, as preliminary experiments showed no performance benefit. Hyperparameters were initially adopted from the original implementations, with additional tuning to stabilize training loss curves. Cross-entropy loss was used for optimization, in combination with the Adam optimizer, and a cosine learning rate scheduler was applied to gradually reduce the learning rate over epochs. Early stopping was applied manually by monitoring the training loss curve to prevent overfitting and promote generalization. Table 2 summarizes the final hyperparameters used for each model.

Several data augmentation techniques were applied exclusively to the training images to enhance generalization and robustness. Gaussian noise (sampled from a normal distribution) and Gaussian blur with a 3 × 3 kernel were each randomly

Table 2. Comparison of model configurations.

Parameter	ResNet50	MViTv2	MambaOut
Parameter count	25.5M	51.4M	84.8M
Image size	224 × 224	224 × 224	224 × 224
Batch size	16	16	16
Epochs	100	100	12
Peak learning rate	$5 \cdot 10^{-4}$	$5 \cdot 10^{-5}$	$5 \cdot 10^{-4}$
Loss function	cross-entropy	cross-entropy	cross-entropy
Optimizer	AdamW	AdamW	AdamW
Learning rate scheduler	Cosine	Cosine	Cosine

applied with a probability of 0.5. Additional augmentations included random horizontal flipping ($p = 0.5$) and random brightness and contrast adjustments scaled between 0.5 and 2×. All training was conducted in Python 3.10 and PyTorch 2.5.1 on an 11GB NVIDIA GeForce GTX 1080 Ti GPU.

2.4 Experiments

In experiment 1, models were trained on a single dataset and evaluated across all test sets to assess cross-dataset generalizability. This quantifies domain overfitting and the feasibility of single-source training.

In experiment 2, a leave-one-dataset-out (LODO) approach was used to train on nine datasets and evaluate on the held-out dataset. To standardize data contributions and limit computational load, training sets were capped at 300 images per dataset (150 benign/150 malignant).

In experiment 3, models were trained on a reduced subset of all datasets (again capped at 300 images per dataset), including the one used for testing, to simulate real-world constraints with limited in-house data. This experiment evaluated whether combining small amounts of diverse external data improves generalizability despite limited application-specific data. Performance was evaluated on the unseen test sets using classification accuracy and the area-under-receiver-operating-characteristics curve (AUROC).

3 Results

3.1 Experiment 1: Cross-Dataset Generalization

Experiment 1 evaluated how well each model generalized across datasets when trained on a single dataset and tested on all others. Resulting accuracy and AUC values are visualized in Figs. 2a–2f and 3. ResNet50 showed strong internal validation (accuracy: 0.82 ± 0.07, AUC: 0.88 ± 0.06), with top results on D1 and D7. However, cross-dataset generalization was limited (accuracy: 0.63 ± 0.09, AUC: 0.69 ± 0.10), with large performance drops on unseen datasets. Training on D1 and D6 yielded the best generalization, while D4 and D8 were least transferable as training sets. D3 and D7 consistently served as favorable test sets, while D4 and D10 were more challenging. MViTv2 demonstrated stronger generalization to other test sets (accuracy: 0.69 ± 0.09, AUC: 0.78 ± 0.10) than ResNet50, particularly when trained on D6. Internal validation was also high (accuracy: 0.85 ± 0.07, AUC: 0.91 ± 0.07). Performance was most robust on D3 as test set and weakest on D10, suggesting D3's data distribution aligns well with various training sets, while D10 presents more challenging conditions for the model. MambaOut performed comparably to MViTv2 in internal validation (accuracy: 0.86 ± 0.04, AUC: 0.93 ± 0.04). Cross-dataset accuracy equaled 0.66 ± 0.08 and AUC 0.77 ± 0.09. D6 again yielded the best generalization, whereas training on D7 and D8 led to poorer cross-dataset results. As with other models, D3 emerged as the most consistently well-performing test set, while D10 posed challenges. In conclusion, MViTv2 exhibited the best overall cross-dataset generalizability, while ResNet50 was the most sensitive to dataset variability.

Test → Train↓	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	Avg ACC
D1	**0,91**	0,70	0,73	0,60	0,64	0,69	0,73	0,71	0,78	0,60	0,71
D2	0,62	**0,70**	0,55	0,52	0,68	0,57	0,60	0,58	0,68	0,55	0,61
D3	0,69	0,65	**0,86**	0,56	0,70	0,59	0,70	0,67	0,64	0,55	0,66
D4	0,53	0,62	0,58	**0,81**	0,53	0,62	0,54	0,59	0,50	0,72	0,60
D5	0,63	0,59	0,66	0,43	**0,88**	0,55	0,64	0,53	0,62	0,54	0,61
D6	0,79	0,69	0,80	0,57	0,77	**0,81**	0,75	0,68	0,68	0,60	0,71
D7	0,49	0,61	0,61	0,57	0,58	0,68	**0,91**	0,59	0,54	0,51	0,61
D8	0,73	0,53	0,73	0,44	0,61	0,53	0,63	**0,77**	0,62	0,45	0,60
D9	0,69	0,60	0,79	0,56	0,69	0,64	0,75	0,78	**0,82**	0,52	0,68
D10	0,66	0,81	0,70	0,66	0,59	0,65	0,74	0,67	0,70	**0,73**	0,69
Avg ACC	0,67	0,65	0,70	0,57	0,67	0,63	0,70	0,66	0,66	0,58	

(a) Accuracy ResNet50

Test → Train↓	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	Avg AUC
D1	**0,95**	0,76	0,84	0,65	0,73	0,75	0,82	0,80	0,83	0,68	0,78
D2	0,65	**0,78**	0,56	0,51	0,72	0,68	0,63	0,63	0,65	0,57	0,64
D3	0,80	0,72	**0,94**	0,62	0,88	0,61	0,78	0,78	0,69	0,54	0,74
D4	0,49	0,71	0,68	**0,90**	0,63	0,79	0,56	0,60	0,54	0,82	0,67
D5	0,63	0,67	0,76	0,48	**0,93**	0,60	0,69	0,56	0,64	0,56	0,65
D6	0,85	0,79	0,89	0,58	0,85	**0,86**	0,80	0,72	0,78	0,65	0,78
D7	0,67	0,76	0,65	0,58	0,57	0,72	**0,96**	0,69	0,63	0,50	0,67
D8	0,77	0,59	0,81	0,51	0,69	0,60	0,67	**0,81**	0,68	0,46	0,66
D9	0,73	0,68	0,88	0,59	0,78	0,68	0,83	0,84	**0,88**	0,63	0,75
D10	0,70	0,88	0,77	0,75	0,73	0,72	0,80	0,74	0,78	**0,82**	0,77
Avg AUC	0,72	0,73	0,78	0,62	0,75	0,70	0,75	0,72	0,71	0,62	

(b) AUC ResNet50

Test → Train↓	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	Avg ACC
D1	**0,96**	0,70	0,87	0,62	0,70	0,62	0,81	0,81	0,70	0,70	0,75
D2	0,69	**0,80**	0,72	0,72	0,57	0,71	0,75	0,76	0,76	0,77	0,73
D3	0,79	0,76	**0,91**	0,68	0,82	0,70	0,68	0,75	0,70	0,62	0,74
D4	0,60	0,72	0,64	**0,92**	0,56	0,67	0,74	0,58	0,66	0,76	0,69
D5	0,63	0,55	0,65	0,38	**0,86**	0,68	0,63	0,54	0,68	0,48	0,61
D6	0,86	0,80	0,86	0,69	0,77	**0,81**	0,82	0,77	0,80	0,65	0,78
D7	0,57	0,65	0,70	0,67	0,66	0,73	**0,85**	0,69	0,62	0,62	0,68
D8	0,76	0,68	0,84	0,54	0,65	0,58	0,65	**0,84**	0,70	0,54	0,68
D9	0,71	0,72	0,81	0,64	0,67	0,61	0,76	0,78	**0,84**	0,53	0,71
D10	0,71	0,72	0,82	0,77	0,57	0,66	0,82	0,71	0,74	**0,72**	0,72
Avg ACC	0,73	0,71	0,78	0,66	0,68	0,68	0,75	0,72	0,72	0,64	

(c) Accuracy MViTv2

Test → Train↓	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	Avg AUC
D1	**0,99**	0,83	0,95	0,73	0,77	0,75	0,90	0,88	0,80	0,76	0,84
D2	0,80	**0,86**	0,92	0,79	0,70	0,79	0,84	0,80	0,83	0,80	0,81
D3	0,86	0,84	**1,00**	0,71	0,87	0,75	0,81	0,84	0,88	0,63	0,82
D4	0,71	0,79	0,81	**0,97**	0,67	0,78	0,93	0,71	0,85	0,89	0,81
D5	0,72	0,57	0,88	0,42	**0,95**	0,75	0,65	0,61	0,72	0,53	0,68
D6	0,92	0,84	0,92	0,73	0,82	**0,88**	0,90	0,82	0,91	0,75	0,85
D7	0,63	0,82	0,78	0,74	0,71	0,80	**0,93**	0,78	0,76	0,66	0,76
D8	0,83	0,76	0,94	0,60	0,69	0,68	0,75	**0,90**	0,76	0,64	0,76
D9	0,74	0,79	0,92	0,65	0,74	0,64	0,89	0,84	**0,88**	0,55	0,76
D10	0,80	0,81	0,91	0,87	0,73	0,78	0,94	0,74	0,85	**0,78**	0,82
Avg AUC	0,80	0,79	0,90	0,72	0,77	0,76	0,85	0,79	0,82	0,70	

(d) AUC MViTv2

Test → Train↓	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	Avg ACC
D1	**0,93**	0,65	0,79	0,61	0,72	0,70	0,73	0,78	0,76	0,73	0,74
D2	0,73	**0,83**	0,73	0,72	0,54	0,66	0,71	0,70	0,66	0,74	0,70
D3	0,58	0,61	**0,89**	0,57	0,71	0,67	0,63	0,63	0,64	0,57	0,65
D4	0,58	0,67	0,64	**0,89**	0,56	0,61	0,59	0,65	0,60	0,80	0,66
D5	0,65	0,72	0,70	0,52	**0,89**	0,69	0,70	0,62	0,68	0,54	0,67
D6	0,78	0,76	0,85	0,71	0,69	**0,87**	0,80	0,78	0,76	0,65	0,77
D7	0,51	0,53	0,57	0,62	0,60	0,53	**0,83**	0,56	0,50	0,54	0,58
D8	0,73	0,58	0,79	0,56	0,60	0,59	0,71	**0,84**	0,70	0,49	0,66
D9	0,69	0,75	0,70	0,57	0,69	0,60	0,63	0,79	**0,82**	0,56	0,68
D10	0,65	0,77	0,77	0,57	0,57	0,72	0,72	0,76	0,70	**0,80**	0,72
Avg ACC	0,68	0,69	0,74	0,65	0,66	0,66	0,71	0,71	0,68	0,64	

(e) Accuracy MambaOut

Test → Train↓	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	Avg AUC
D1	**0,98**	0,78	0,89	0,74	0,75	0,82	0,88	0,84	0,76	0,79	0,82
D2	0,79	**0,88**	0,88	0,76	0,72	0,74	0,81	0,79	0,76	0,79	0,79
D3	0,74	0,79	**0,96**	0,61	0,82	0,79	0,79	0,82	0,84	0,62	0,78
D4	0,78	0,84	0,86	**0,95**	0,69	0,77	0,74	0,69	0,78	0,88	0,80
D5	0,77	0,82	0,84	0,60	**0,98**	0,80	0,83	0,68	0,76	0,63	0,77
D6	0,81	0,85	0,94	0,77	0,77	**0,91**	0,91	0,82	0,86	0,77	0,84
D7	0,59	0,68	0,73	0,75	0,68	0,80	**0,92**	0,78	0,73	0,66	0,73
D8	0,78	0,68	0,89	0,57	0,65	0,63	0,80	**0,91**	0,78	0,54	0,72
D9	0,74	0,79	0,88	0,56	0,75	0,64	0,80	0,86	**0,93**	0,48	0,74
D10	0,78	0,88	0,85	0,85	0,77	0,78	0,93	0,83	0,82	**0,84**	0,83
Avg AUC	0,78	0,80	0,87	0,72	0,76	0,77	0,84	0,80	0,80	0,70	

(f) AUC MambaOut

Fig. 2. Overview of heat maps for accuracy (left) and AUC (right) in Experiment 1 across the three different models. Green represents the highest values and red the lowest. Bold values indicate internal validation results. (Color figure online)

3.2 Experiment 2: Leave-One-Dataset-Out Generalization

LODO experiments tested each model's ability to generalize when trained on all but one dataset. Table 3 summarizes the LODO performance of each model. MViTv2 achieved the highest average classification accuracy (0.77 ± 0.06) and AUC (0.85 ± 0.07), followed by MambaOut (0.75 ± 0.05 and 0.84 ± 0.06) and ResNet50 (0.74 ± 0.04 and 0.84 ± 0.04). All models performed well on D3 and D7 and consistently poorly on D5 (especially for ResNet50), indicating that D3 resembles the aggregate training distribution while D5 is more divergent.

3.3 Experiment 3: Diverse External Data with Limited In-domain

Lastly, the added value of combining small amounts of diverse external data with limited application-specific data for generalizability was evaluated and summarized in Table 4 for every model. MViTv2 achieved the highest average accuracy across combinations (0.81 ± 0.06), while MambaOut achieved the highest AUC (0.89 ± 0.04). ResNet50 achieved lower average performance (accuracy of 0.77 ± 0.04 and AUC of 0.85 ± 0.05). The best results were achieved when trained on D1, D3, and D7. The relatively stable performance across datasets supports the models' generalization capabilities, even when trained with just a fraction

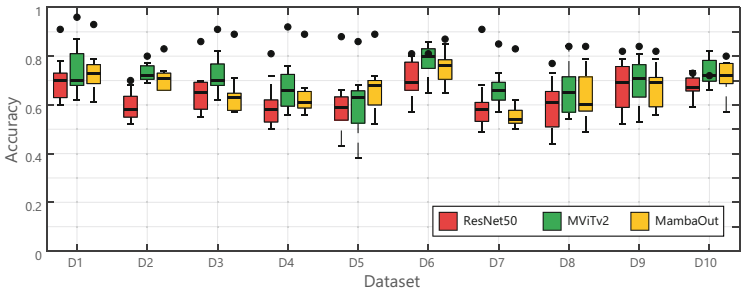


Fig. 3. Boxplots showing the ranges of cross-dataset validation accuracy in Experiment 1 across the different datasets. The black circles represent the corresponding internal validation accuracies.

Table 3. Accuracy and AUC results for the Leave-One-Dataset-Out experiment across the three models (Experiment 2).

Test set	Accuracy				AUC			
	ResNet50	MViTv2	MambaOut	Mean±Std	ResNet50	MViTv2	MambaOut	Mean±Std
D1	0.78	0.80	0.78	0.79 ± 0.01	0.86	0.93	0.89	0.89 ± 0.04
D2	0.76	0.74	0.72	0.74 ± 0.02	0.84	0.79	0.79	0.81 ± 0.03
D3	0.78	0.86	0.87	0.84 ± 0.05	0.91	0.94	0.96	0.94 ± 0.03
D4	0.74	0.77	0.73	0.75 ± 0.02	0.84	0.82	0.80	0.82 ± 0.02
D5	0.64	0.75	0.69	0.69 ± 0.06	0.76	0.80	0.76	0.77 ± 0.02
D6	0.70	0.69	0.69	0.69 ± 0.01	0.83	0.78	0.86	0.82 ± 0.04
D7	0.77	0.86	0.80	0.81 ± 0.05	0.89	0.95	0.89	0.91 ± 0.03
D8	0.75	0.75	0.74	0.75 ± 0.01	0.79	0.83	0.80	0.81 ± 0.02
D9	0.72	0.80	0.74	0.75 ± 0.04	0.82	0.90	0.85	0.86 ± 0.04
D10	0.74	0.71	0.74	0.73 ± 0.02	0.81	0.79	0.79	0.80 ± 0.01
Mean±Std	0.74 ± 0.04	0.77 ± 0.06	0.75 ± 0.05		0.84 ± 0.04	0.85 ± 0.07	0.84 ± 0.06	

Table 4. Accuracy and AUC results for the deep learning models tested on a small subset of all datasets (Experiment 3).

Test set	Accuracy				AUC			
	ResNet50	MViTv2	MambaOut	Mean ± Std	ResNet50	MViTv2	MambaOut	Mean ± Std
D1	0.84	0.88	0.85	0.86 ± 0.02	0.93	0.94	0.90	0.92 ± 0.02
D2	0.75	0.77	0.81	0.78 ± 0.03	0.84	0.83	0.86	0.84 ± 0.02
D3	0.81	0.86	0.89	0.85 ± 0.04	0.91	0.96	0.94	0.94 ± 0.03
D4	0.71	0.79	0.82	0.77 ± 0.06	0.79	0.84	0.90	0.84 ± 0.06
D5	0.80	0.81	0.74	0.78 ± 0.04	0.82	0.88	0.87	0.86 ± 0.03
D6	0.73	0.69	0.67	0.70 ± 0.03	0.82	0.79	0.81	0.81 ± 0.02
D7	0.76	0.91	0.87	0.85 ± 0.08	0.92	0.96	0.97	0.95 ± 0.03
D8	0.79	0.83	0.77	0.80 ± 0.03	0.84	0.90	0.88	0.87 ± 0.03
D9	0.76	0.80	0.76	0.77 ± 0.02	0.83	0.85	0.89	0.86 ± 0.03
D10	0.71	0.80	0.81	0.77 ± 0.06	0.83	0.87	0.86	0.85 ± 0.02
Mean ± Std	0.77 ± 0.04	0.81 ± 0.06	0.80 ± 0.07		0.85 ± 0.05	0.88 ± 0.06	0.89 ± 0.04	

of the available data. Overall accuracy improved compared to training on single datasets (Experiment 1) and was comparable to LODO results (Experiment 2).

4 Discussion

Our results demonstrate that model performance for BUS lesion classification is highly sensitive to the composition and diversity of the training data. In particular, models trained on single-source datasets performed well on their internal validation set but exhibited significant performance drops on external test sets, confirming the presence of domain overfitting. This effect was most pronounced for ResNet50, but also observed for the transformer-based MViTv2 and state-space-based MambaOut models.

Our experiments showed that even relatively simple models achieve better generalization and robustness when trained on small but diverse datasets (varying in scanner type, operator, demographics, or protocol) than on large homogeneous datasets. This underscores that data diversity contributes more to performance than sheer volume. This has important implications for clinical AI development, especially in low-data or resource-constrained hospitals: sharing small, anonymized datasets across institutions may enable more generalizable AI tools without requiring full data harmonization or centralization. Notably, the MViTv2 architecture seems to achieve the most consistent performance across datasets, suggesting that ViTs may better handle cross-domain variation, possibly due to their attention mechanisms capturing more global features. Interestingly, the MambaOut model showed potential, but did not consistently outperform either CNNs or ViTs. This aligns with earlier work noting its strong performance on natural images, but suggests its application to clinical BUS tasks still requires architectural tuning or domain-specific pretraining. Its high parameter efficiency and state-space formulation remain promising for future use in low-resource or mobile settings.

Compared to prior studies, our internal validation results (Experiment 1) for MViTv2, showing accuracy values of up to 0.96 and AUC values up to 1.00, are in line with the best-performing CNN-based models reported in the literature. However, unlike most existing work, we also evaluated generalization across unseen datasets, where performance dropped, highlighting the challenges of cross-domain robustness. While Kalafi et al. [9] and Gu et al. [8] reported accuracies up to 0.93 using multi-source training data, their datasets were more homogeneous, often from the same acquisition protocol or fewer institutions. In contrast, our setup (Experiment 3) involved greater scanner, protocol, and population diversity, reflecting a more realistic clinical scenario, but also increasing generalization difficulty with resulting accuracies ranging from 0.69 to 0.91.

However, some limitations remain. First, public datasets differ in image and annotation quality and labeling protocols, which may introduce bias. Second, although ten datasets were included, the absence of true individual demographic metadata limits the ability to link model robustness to patient-specific factors. Future work should explore harmonized multi-center datasets with richer metadata to systematically assess model fairness and bias. Third, although we tested three model families, we did not perform architecture-level ablation experiments. Lastly, in the clinical setting, minimizing false negatives is crucial to avoid missing malignant lesions. While our study prioritized overall accuracy and AUC,

future work should shift focus toward maximizing sensitivity. The resulting specificity can then be evaluated to estimate the number of unnecessary biopsies of benign lesions that can be avoided using BUS screening.

5 Conclusion

This study highlights the key role of training dataset diversity in developing robust and generalizable AI models for breast ultrasound lesion classification. Models trained on diverse, multi-source datasets consistently outperformed those trained on single datasets, with vision transformers demonstrating the best cross-domain robustness. Our results caution against deploying AI models trained on limited datasets, as such models may not generalize well to real-world clinical data and could perpetuate healthcare disparities. Future development should prioritize training on heterogeneous datasets, in terms of patient ethnicity and ultrasound hardware, and include performance reporting across diverse test cohorts to ensure fairness and reliability in clinical use.

References

1. World Health Organization. Breast cancer—who.int (2024). https://www.who.int/news-room/fact-sheets/detail/breast-cancer. Accessed 05 Dec 2024
2. Hofvind, S., Ursin, G., Tretli, S., Sebuødegård, S., Møller, B.: Breast cancer mortality in participants of the Norwegian breast cancer screening program. Cancer **119**(17), 3106–3112 (2013)
3. Cheng, H.-D., Shan, J., Wen, J., Guo, Y., Zhang, L.: Automated breast cancer detection and classification using ultrasound images: A survey. Pattern Recogn. **43**(1), 299–317 (2010)
4. Alruily, M., Mahmoud, A.A., Allahem, H., Mostafa, A.M., Shabana, H., Ezz, M.: Enhancing breast cancer detection in ultrasound images: An innovative approach using progressive fine-tuning of vision transformer models. Int. J. Intell. Syst. **2024**(1), 6528752 (2024)
5. Cao, Z., Duan, L., Yang, G., Yue, T., Chen, Q.: An experimental study on breast lesion detection and classification from ultrasound images using deep learning architectures. BMC Med. Imaging **19**, 1–9 (2019)
6. Ali, M.D., et al.: Breast cancer classification through meta-learning ensemble technique using convolution neural networks. Diagnostics **13**(13), 2242 (2023)
7. Marco, P., Ricciardi, V., Montesano, M., Cassano, E., Origgi, D.: Transfer learning classification of suspicious lesions on breast ultrasound: is there room to avoid biopsies of benign lesions? Eur. Radiol. Exp. **8**(1), 121 (2024)
8. Yang, G., et al.: Ultrasound-based deep learning in the establishment of a breast lesion risk stratification system: a multicenter study. Eur. Radiol. **33**(4), 2954–2964 (2023)
9. Kalafi, E.Y., et al.: Classification of breast cancer lesions in ultrasound images by using attention layer and loss ensemble in deep convolutional neural networks. Diagnostics **11**(10), 1859 (2021)
10. Han, S., et al.: A deep learning framework for supporting the classification of breast lesions in ultrasound images. Phys. Med. Biol. **62**(19), 7714 (2017)

11. Eroğlu, Y., Yildirim, M., Çinar, A.: Convolutional neural networks based classification of breast ultrasonography images by hybrid method with respect to benign, malignant, and normal using mrmr. Comput. Biol. Med. **133**, 104407 (2021)
12. Mo, Y., et al.: Hover-trans: Anatomy-aware hover-transformer for ROI-free breast cancer diagnosis in ultrasound images. IEEE Trans. Med. Imaging **42**(6), 1696–1706 (2023)
13. Gheflati, B., Rivaz, H.: Vision transformers for classification of breast ultrasound images. In: 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), pp. 480–483. IEEE (2022)
14. Yu, W., Wang, X.: MambaOut: do we really need mamba for vision? arXiv preprint arXiv:2405.07992 (2024)
15. Nasiri-Sarvi, A., Hosseini, M.S., Rivaz, H.: Vision mamba for classification of breast ultrasound images. arXiv preprint arXiv:2407.03552 (2024)
16. American Cancer Society. Breast cancer facts & figures 2024-2025 (2024). Accessed 09 Dec 2024
17. Gordon, P.B.: The impact of dense breasts on the stage of breast cancer at diagnosis: a review and options for supplemental screening. Curr. Oncol. **29**(5), 3595–3636 (2022)
18. Koo, C., et al.: Validating racial and ethnic non-bias of artificial intelligence decision support for diagnostic breast ultrasound evaluation. J. Med. Imaging **10**(6), 061108 (2023)
19. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
20. Li, Y., et al.: MViTv2: improved multiscale vision transformers for classification and detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4804–4814 (2022)
21. Ardakani, A.A., Mohammadi, A., Mirza-Aghazadeh-Attari, M., Acharya, U.R.: An open-access breast lesion ultrasound image database: applicable in artificial intelligence studies. Comput. Biol. Med. **152**, 106438 (2023)
22. Al-Dhabyani, W., Gomaa, M., Khaled, H., Fahmy, A.: Dataset of breast ultrasound images. Data Brief **28**, 104863 (2020)
23. Iqbal, A., Sharif, M.: Memory-efficient transformer network with feature fusion for breast tumor segmentation and classification task. Eng. Appl. Artif. Intell. **127**, 107292 (2024)
24. Lin, Z., Lin, J., Zhu, L., Fu, H., Qin, J., Wang, L.: A new dataset and a baseline model for breast lesion detection in ultrasound videos. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022. LNCS, vol. 13433, pp. 614–623. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16437-8_59
25. Gómez-Flores, W., Gregorio-Calas, M.J., Albuquerque Pereira, W.C.: BUS-BRA: a breast ultrasound dataset for assessing computer-aided diagnosis systems. Med. Phys. **51**(4), 3110–3123 (2024)
26. Vallez, N., Bueno, G., Deniz, O., Rienda, M.A., Pastor, C.: BUS-UCLM: breast ultrasound lesion segmentation dataset (2024)
27. Rodtook, A., Kirimasthong, K., Lohitvisate, W., Makhanov, S.S.: Automatic initialization of active contours and level set method in ultrasound images of breast abnormalities. Pattern Recogn. **79**, 172–182 (2018)
28. Yap, M.H., et al.: Automated breast ultrasound lesions detection using convolutional neural networks. IEEE J. Biomed. Health Inform. **22**(4), 1218–1226 (2017)
29. Geertsma, T.: Ultrasound cases, 5 breast and axilla. https://www.ultrasoundcases.info/cases/breast-and-axilla/. Accessed 29 Jan 2025

FedBC: Privacy-Preserving Breast Cancer Diagnosis from Ultrasound Images Using Federated Learning

Faisal Ahmed[1,2(✉)], David Sánchez[1], Josep Domingo-Ferrer[1], and Zouhair Haddi[2]

[1] Department of Computer Engineering and Mathematics, CYBERCAT-Center for Cybersecurity Research of Catalonia, Universitat Rovira i Virgili, Tarragona, Catalonia
{david.sanchez,josep.domingo}@urv.cat
[2] NVISION Systems and Technologies SL, Barcelona, Catalonia
{faisal.ahmed,zouhair.haddi}@nvision.es

Abstract. Accurate diagnosis of breast cancer in dense breasts requires expert radiologists to examine multiple ultrasound images per patient. This diagnosis procedure is tedious, time-consuming and prone to misdiagnosis due to human fatigue. AI-aided diagnosis systems can help alleviate this burden. However, vast amounts of data from multiple hospitals, diverse patient demographics, imaging scanners, and protocols are required to develop accurate, robust, and generalizable AI models. Obtaining such a mixture of data is quite challenging due to privacy concerns, data ownership issues, and regulatory constraints. To address this problem, we introduce a privacy-preserving and regulatory-compliant method for breast cancer diagnosis from ultrasound images using federated learning. This allows hospitals to keep their data on their premises while collaboratively training the AI model, only exchanging the model parameters trained on their private data. Experimental evaluation within a real federated data setting shows that our method achieved higher diagnostic accuracy than standardized centralized training while ensuring compliance with data privacy regulations.

Keywords: Breast cancer diagnosis · Federated learning · Data privacy

1 Introduction

Breast cancer causes the highest mortality among women worldwide among all cancer-related deaths. Early diagnosis of breast cancer significantly improves patient outcomes and minimizes the risk of mortality [21]. Ultrasound (US) imaging is a widely used modality for the detection and diagnosis of breast cancer, particularly in resource-constrained settings, due to its cost-effectiveness, non-invasiveness, and ability to provide real-time imaging [20]. However, manual breast cancer diagnosis from US images is time-consuming and requires expert radiologists. In addition, with the increasing number of cases of breast cancer, the

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
T. Zhang et al. (Eds.): Deep-Breath 2025, LNCS 16142, pp. 31–40, 2026.
https://doi.org/10.1007/978-3-032-05559-0_4

volume of US images that radiologists must interpret also increases, which can lead to diagnostic delays and errors. Therefore, a precise AI-assisted diagnosis of breast cancer from US images is essential to support medical decision-making.

Recent advances in deep learning have shown significant success in automating medical image analysis for diagnosis [17,19,23] and prediction [3,6,12]. However, deep learning requires vast amounts of data from multiple hospitals, imaging scanners, and diverse patient demographics to develop accurate, robust, and generalizable models. Such a wealth of data is often difficult to obtain due to privacy concerns, data ownership issues, and regulatory restrictions such as the Health Insurance Portability and Accountability Act (HIPAA) [1] and the EU General Data Protection Regulation (GDPR) [2]. Moreover, processing large volumes of data in a single institution requires a high-performance computing infrastructure, which is costly and often unavailable in resource-constrained settings, such as those in developing countries.

Federated learning (FL) [16] has emerged as a promising paradigm for training collaborative models between multiple institutions without the need to share data directly. By enabling decentralized learning, FL intrinsically preserves data privacy, distributes computational load across multiple institutions or edge devices, and leverages multi-institutional datasets to enhance model generalization. In our setting, FL can help overcome the limitations associated with developing deep learning models for the diagnosis of breast cancer using US images.

To our knowledge, the feasibility of applying FL for breast cancer diagnosis using US images has not yet been explored. Our study aims to address this research gap through the following key contributions:

1. We propose a privacy-preserving and regulatory-compliant framework that leverages FL to diagnose breast cancer from US images. Our framework enhances model generalizability by integrating multi-institutional datasets encompassing diverse patient demographics, imaging protocols, and US scanners from different vendors.
2. We curate multiple data sets that feature variations in imaging protocols, patient demographics, number of images, and US systems, reflecting real-world data heterogeneity between different hospitals.
3. We report empirical results that compare our FL setting with conventional centralized learning (CL).

The remainder of the paper is organized as follows. Section 2 discusses related work on AI-based breast cancer diagnosis. Section 3 describes our method for the training of FL-based breast cancer models. Section 4 details the datasets and experimental setup, while Sect. 5 reports and discusses the results obtained. Conclusions and several lines of future research are described in Sect. 6.

2 Related Work

Several solutions have been proposed to automate the diagnosis of breast cancer from US images. Most of them operate in a centralized setting, where a single computing server aggregates all images from different sources and performs

model training and validation. Most of these solutions follow a common workflow comprising three steps: *image preprocessing*, *lesion segmentation*, and *lesion classification*.

Preprocessing techniques include: (1) *data augmentation* to mitigate class imbalance and improve model generalization; (2) *simple linear iterative clustering* [11] to address complex artifacts and noise; (3) *anisotropic diffusion* [8], *interference-based despeckling* [8], and *bilateral filtering* [24] to reduce noise and speckle noise while preserving edges and fine details; (4) *morphological opening and closing operations* [24] on the original image using the binary mask to refine tumor boundaries; (5) *sigmoid filtering* [8] and *fuzzy enhancement* [24] to improve image quality and contrast, making lesions more distinguishable from surrounding tissues. For lesion segmentation, related works use U-Net [15], fully convolutional networks [14] or UNet variants [11,22]. For classification, [11] and [15] adopted a deep transfer learning approach, with the former fine-tuning a pre-trained VGG19 model and the latter fine-tuning a pre-trained ResNet50 model. On the other hand, three studies [8,14,22] train a wavelet neural network, a custom convolutional neural network (CNN), and a CNN with Inception blocks and an attention mechanism, respectively, from scratch. Few studies have also explored traditional machine learning techniques. [8] utilized *wavelet neural networks*, along with manually extracted morphological and texture features. Stacking ensembles of multi-layer perceptrons and support vector machines, combined with transfer learning features of segmented lesions obtained from pre-trained VGG16, VGG19, and InceptionV3 models, were used in [8].

The works discussed above have several limitations: (1) They require aggregating patients' US images in a central computing platform from multiple data sources (*i.e.*, hospitals), potentially compromising patient privacy and violating regulatory compliance requirements such as HIPAA, and GDPR; (2) The models are trained and validated on single or limited source data, and their performance across diverse or out-of-domain datasets and varying imaging conditions has not been extensively evaluated; (3) They rely on multiple models for segmentation and classification along with extensive preprocessing, which is computationally expensive and potentially limits the pipeline's applicability in resource-constrained environments and edge devices.

3 Methods

Figure 1 depicts the privacy-preserving training setup for our FL-based breast cancer diagnosis model. Each hospital retains its private dataset of breast US images, corresponding masks, and labels. They perform model updates locally and share the updated model parameters with a server, which coordinates the model's training across all hospitals over several communication rounds.

In the communication round t, the server distributes the global model parameters θ_t to all participating hospitals. Each hospital initializes its local model with θ_t as a starting point and updates it using its private data, producing local models with parameters $\theta_t^1, \theta_t^2, \ldots, \theta_t^K$. These updated parameters are sent

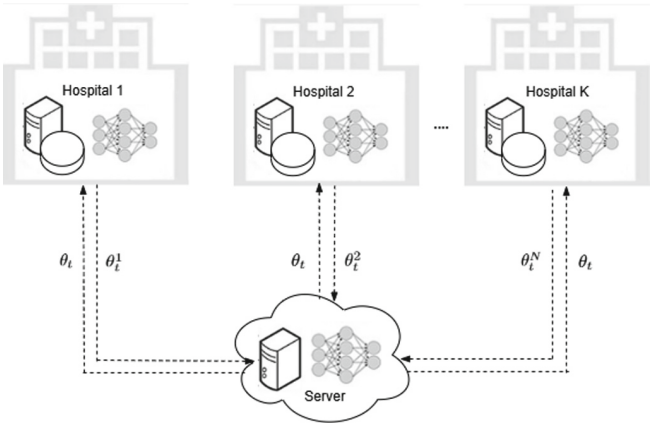


Fig. 1. FL framework for privacy-preserving breast cancer diagnosis using US images

back to the server, which aggregates them to refine the global model. This process repeats over multiple rounds until convergence. Only model updates are exchanged between the server and hospitals, thereby ensuring that no raw data are shared.

Model Architecture for Breast Cancer Diagnosis. We utilized the ResNet-18 [10] architecture as a breast cancer prediction model. The input layer of the network was modified to handle a two-channel input $\mathbf{x} \in \mathbb{R}^{H \times W \times 2}$ –consisting of a grayscale breast US image and its corresponding mask, both with height H and width W. In a similar manner, the final layer of the network was restructured to perform binary classification by replacing it with a single output neuron. This neuron employs a sigmoid activation function, which maps the network's output to a probability value between 0 and 1, enabling the model to distinguish between two classes –*benign* and *malignant*– based on the predicted probability.

Model Parameter Update at Hospital Premises. The local training objective of hospital k at training round t is to update the global model parameter θ_t, which maps input data X to target outputs Y in the hospital's private dataset while minimizing a given loss function $\mathcal{L}(\theta_t)$. Formally, for the private dataset $D_k = \{(x_i, y_i)\}_{i=1}^{n^k}$ of hospital k, the optimization objective is

$$\theta_t^k = \arg\min_{\theta_t} \mathbb{E}_{(x,y)\sim D_k} \mathcal{L}(\theta_t(x), y), \tag{1}$$

where $\theta_t(x)$ is the predicted output of the model, y is the ground truth label, $\mathcal{L}$ is the loss function and $\mathbb{E}$ denotes the expectation over the distribution of the data of hospital k.

Since breast cancer diagnosis is a binary classification problem and the model architecture employs a sigmoid activation function in the classification layer, we use binary cross-entropy as the loss function $\mathcal{L}$. For the dataset of hospital k

with n_k samples, where $\mathbf{x}_i$ is the i-th input and $y_i \in \{0, 1\}$ is its corresponding label, the cross-entropy loss for a single sample, given model parameter θ_t, is:

$$\mathcal{L}_{\text{CE}}(\theta_t, \mathbf{x}_i, y_i) = -\left[y_i \log p_i + (1 - y_i) \log(1 - p_i)\right], \tag{2}$$

where $p_i = \sigma(\theta_t^T \mathbf{x}_i)$, and $\sigma(z) = \frac{1}{1+e^{-z}}$ is the sigmoid function.

To prevent overfitting, the loss function incorporates an L2 regularization term that penalizes large parameter values. The final loss function over n_k samples of hospital k combines the classification loss and the regularization term, defined as

$$\mathcal{L}(\theta_t, \mathbf{x}_i, y_i) = \underbrace{\frac{1}{n_k} \sum_{i=1}^{n_k} \mathcal{L}_{\text{CE}}(\theta_t, \mathbf{x}_i, y_i)}_{\text{classification loss}} + \underbrace{\frac{\lambda}{2} \|\theta_t\|_2^2}_{\text{regularization term}}, \tag{3}$$

where λ controls the regularization strength.

The loss $\mathcal{L}$ is minimized, and the model parameter is updated using stochastic gradient descent as follows:

$$\theta_t^k = \theta_t - \eta \nabla \mathcal{L}(\theta_t; D_k), \tag{4}$$

where $\mathcal{L}(\theta_t; D_k)$ is the regularized binary cross-entropy loss function in Eq. (3), $\nabla \mathcal{L}(\theta_t; D_k)$ is the gradient computed on the local dataset D_k, and η is the local learning rate.

Aggregation of Parameters at the Server. After performing multiple local updates within their premises during training round t, hospitals transmit their updated model parameters, θ_t^k, to the server. The server then aggregates these parameters to compute the global model update for the federated training round $t + 1$ using the Federated Averaging (FedAvg) [16] algorithm as follows:

$$\theta_{t+1} = \sum_{k=1}^{K} \frac{n_k}{n} \theta_t^k, \tag{5}$$

where n_k represents the number of data samples held by hospital k, and n denotes the total number of samples across all participating hospitals.

4 Experimental Setup

Datasets. We curated three datasets from public sources: *BUS-BRA* [9], *BUSI* [5], and *QAMEBI* [7]. These datasets were collected from hospitals in various regions of the world, using different US systems at different times and under varying image acquisition protocols. We further partitioned the datasets into federated hospitals or domains based on ultrasound system specifications, resulting in six distinct groups, referred to as hospitals A through F in Table 1.

All the images were preprocessed using a standard transformation pipeline. Each image was first resized to a fixed resolution of 128×128 pixels to ensure uniform input dimensions. The intensity values of the resized images were then normalized to the range $[-1, 1]$, using a mean of 0.5 and a standard deviation of 0.5.

Table 1. Features of the federated datasets

Hospital	Source	US Machine	Frequency (MHz)	# Patients	# Images
A	BUS-BRA [9]	Toshiba Aplio 300	12–14	72	902
B	BUS-BRA [9]	GE Logiq 5	10–12	490	1064
C	BUS-BRA [9]	GE Logiq 7	10–14	488	809
D	BUS-BRA [9]	GE U-Systems	12–14	14	25
E	BUSI [5]	LOGIQ E & LOGIQ E9 Agile	1–5	–	697
F	QAMEBI [7]	AirPlorer Ultimate	5–18	–	232

Implementation Details and Evaluation Metrics. Model training was conducted at each hospital site using a mini-batch gradient descent optimizer with a batch size of 64 image samples. The initial learning rate was set at 0.001 and a StepLR scheduler was used to adjust the learning rate during training. Each hospital performed local training for 10 epochs per communication round, and FL was executed over 20 communication rounds.

The hardware setup consisted of a remote server equipped with an Intel Xeon processor, 64GB of RAM, and an NVIDIA A10 GPU with 24 GB of VRAM. The software environment included the Ubuntu 22.04.3 LTS operating system, CUDA 12.2, and NVIDIA driver version 570.124.06. Models were implemented and trained using PyTorch 1.9.0, while Python 3.10.12 was used to develop the FL pipelines. We employed five classification metrics for performance evaluation: *accuracy*, *precision*, *recall*, *F1-score*, and the *area under the ROC curve* (auROC).

5 Results and Discussion

Firstly, we investigate how FL enables each hospital to develop a more generalized model that is specifically tailored to its local data distribution while benefiting from the collective knowledge of the data of the other participating hospitals. To demonstrate this benefit, we compare the performance of individually and independently trained models on each hospital's data with that of a global model jointly trained by all hospitals using FL and evaluated on each hospital's test data. More specifically, we split each hospital's dataset into training and testing sets with an 80:20 ratio. An individual model is trained solely on each hospital's training set and evaluated on its corresponding test set. On the other hand, FL is performed collaboratively by using the training sets from

all hospitals. The resulting global model is then individually evaluated in each hospital's test set.

Table 2. FL vs individual training performance across hospitals. Best figures in bold.

Hospital	Training type	Accuracy↑	Precision↑	Recall↑	F1-Score↑	auROC↑
A	Federated	**89.29**	**0.85**	**0.92**	**0.88**	**0.91**
	Individual	82.72	0.73	0.64	0.68	0.86
B	Federated	**89.51**	**0.86**	**0.77**	**0.81**	**0.93**
	Individual	82.72	0.73	0.64	0.68	0.86
C	Federated	**81.22**	0.70	**0.85**	**0.76**	0.88
	Individual	79.56	**0.82**	0.55	0.66	**0.90**
D	Federated	**100.00**	**1.00**	**1.00**	**1.00**	**1.00**
	Individual	60.00	0.50	0.50	0.50	0.67
E	Federated	**97.69**	**1.00**	**0.93**	**0.96**	0.98
	Individual	96.95	1.00	0.88	0.94	**1.00**

The results of this evaluation are presented in Table 2, and show that FL consistently outperforms individual training on all metrics in all hospitals. This clearly indicates that federated training improves generalization and robustness to hospital-specific data distributions. This result could be expected, as each hospital trains a model solely on its local data in the individual training setting, which is limited in both size and diversity and may lead to underfitting, overfitting, or poor generalization. In contrast, FL allows hospitals to learn from each other's data while preserving their privacy. This enables training on a larger, more diverse dataset than any single hospital could access, improving parameter estimation and reducing prediction variance. In this sense, model aggregation acts like an ensemble, further boosting model generalization. Together, these factors enhance the model's performance on hospital-specific data distributions.

Secondly, we compare the leave-one-domain-out validation results of FL with those of CL. For this experiment, data from five out of six domains (A to E) are used to simulate five federated hospitals, while the remaining domain (F) serves as an unseen test set to evaluate the global model produced by FL. To enable a fair comparison with CL, data from the same five domains are aggregated into a single combined dataset to train a standalone model. The held-out domain is then used exclusively for evaluation, mirroring the setup used in the FL scenario. This process is repeated six times, with each domain taking a turn as the unseen test set. The remaining five domains are used to train the same model in both FL and CL settings. After all six iterations, performance metrics (*e.g.*, accuracy, precision, recall, F1-score and auROC) are averaged to provide a robust estimate of the generalization performance of each approach. Table 3 reports the results on six unseen sites in terms of five metrics.

The FL approach achieves comparable CL performance on average across all evaluation metrics while offering significant additional benefits: it enhances data privacy, eliminates the demanding need to aggregate data in a central repository, and reduces the demand for highly centralized computing resources by distributing the computational load for model training across participating hospitals. In general, these results demonstrate that FL is a viable solution for developing deep learning models for breast cancer diagnosis in US images.

Table 3. Performance metrics on held-out of domains. Best average figures in bold.

Unseen domain	Setting	Accuracy↑	Precision↑	Recall↑	F1-Score↑	auROC↑
A	FL	82.73	0.81	0.85	0.83	0.91
	CL	80.58	0.83	0.80	0.80	0.92
B	FL	87.76	0.87	0.70	0.77	0.93
	CL	86.53	0.86	0.80	0.82	0.92
C	FL	80.82	0.67	0.78	0.72	0.88
	CL	76.27	0.76	0.79	0.75	0.88
D	FL	100.00	1.00	1.00	1.00	1.00
	CL	100.00	1.00	1.00	1.00	1.00
E	FL	89.95	0.83	0.87	0.85	0.96
	CL	89.34	0.88	0.87	0.87	0.94
F	FL	78.88	0.78	0.83	0.81	0.85
	CL	78.02	0.78	0.77	0.77	0.84
Average	FL	**86.69**	0.83	**0.84**	**0.83**	**0.92**
	CL	85.12	**0.85**	**0.84**	**0.83**	**0.92**

6 Conclusions

We proposed a privacy-preserving and regulatory-compliant framework that leverages FL for breast cancer diagnosis from US images guided by lesion masks. Empirical results and comparative analyses against traditional CL and individual training approaches demonstrate the framework's ability to support multi-institutional collaboration in training a highly generalized model for breast cancer diagnosis. FL also overcomes the reliance on a single high-power computing platform, making it well suited for resource-constrained settings.

In future research, we will extend this framework to address data heterogeneity [13] in different hospitals caused by variations in image acquisition protocols, US systems, patient demographics, and number of images. We also plan to incorporate the MemberShield [4] method to further enhance the robustness of the framework against membership inference attacks [18].

Acknowledgments.. Partial support for this work has been received from Marie Skodowska-Curie Doctoral Networks Actions (HORIZON-MSCA-2021-DN-01-01; 101073222), the Government of Catalonia (ICREA Acadèmia Prizes to J. Domingo-Ferrer and to D. Sánchez, and grant 2021SGR-00115), MCIN/AEI/10.13039/501100011033 and "ERDF A way of making Europe" under grant PID2021-123637NB-I00 "CURLING", and the EU's NextGenerationEU/PRTR via INCIBE (project "HERMES" and INCIBE-URV cybersecurity chair).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Health insurance portability and accountability act of 1996 (1996)
2. Regulation (EU) 2016/679 of the European parliament and of the council on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (2016)
3. Afshar, P., Mohammadi, A., Plataniotis, K.N., Oikonomou, A., Benali, H.: From handcrafted to deep-learning-based cancer radiomics: challenges and opportunities. IEEE Signal Process. Mag. **36**(4), 132–160 (2019)
4. Ahmed, F., Sánchez, D., Haddi, Z., Domingo-Ferrer, J.: MemberShield: a framework for federated learning with membership privacy. Neural Netw. **181**, 106768 (2025)
5. Al-Dhabyani, W., Gomaa, M., Khaled, H., Fahmy, A.: Dataset of breast ultrasound images. Data Brief **28**, 104863 (2020)
6. Allugunti, V.R.: Breast cancer detection based on thermographic images using machine learning and deep learning algorithms. Int. J. Eng. Comput. Sci. **4**(1), 49–56 (2022)
7. Ardakani, A.A., Mohammadi, A., Mirza-Aghazadeh-Attari, M., Acharya, U.R.: An open-access breast lesion ultrasound image database: applicable in artificial intelligence studies. Comput. Biol. Med. **152**, 106438 (2023)
8. Bourouis, S., Band, S.S., Mosavi, A., Agrawal, S., Hamdi, M.: Meta-heuristic algorithm-tuned neural network for breast cancer diagnosis using ultrasound images. Front. Oncol. **12**, 834028 (2022)
9. Gómez-Flores, W., Gregorio-Calas, M.J., de Albuquerque Pereira, W.C.: BUS-BRA: a breast ultrasound dataset for assessing computer-aided diagnosis systems. Med. Phys. **51**(4), 3110–3123 (2024)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference On Computer Vision and Pattern Recognition, pp. 770–778 (2016)
11. Inan, M.S.K., Alam, F.I., Hasan, R.: Deep integrated pipeline of segmentation guided classification of breast cancer from ultrasound images. Biomed. Signal Process. Control **75**, 103553 (2022)
12. Khan, S., Islam, N., Jan, Z., Din, I.U., Rodrigues, J.J.C.: A novel deep learning based framework for the detection and classification of breast cancer using transfer learning. Pattern Recogn. Lett. **125**, 1–6 (2019)
13. Li, Q., Diao, Y., Chen, Q., He, B.: Federated learning on non-IID data silos: an experimental study. In: 2022 IEEE 38th International Conference on Data Engineering (ICDE), pp. 965–978. IEEE (2022)

14. Long, B., Guan, Y., Holden, M.: A two-stage neural network model for breast ultrasound image classification. In: 2023 IEEE 23rd International Conference on Bioinformatics and Bioengineering (BIBE), pp. 129–133. IEEE (2023)
15. Luo, Y., Huang, Q., Li, X.: Segmentation information with attention integration for classification of breast tumor in ultrasound image. Pattern Recogn. **124**, 108427 (2022)
16. McMahan, B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A.: Communication-efficient learning of deep networks from decentralized data. In: Artificial Intelligence and Statistics, pp. 1273–1282. PMLR (2017)
17. Sethi, R., Mehrotra, M., Sethi, D.: Deep learning based diagnosis recommendation for Covid-19 using chest X-rays images. In: 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA), pp. 1–4. IEEE (2020)
18. Shokri, R., Stronati, M., Song, C., Shmatikov, V.: Membership inference attacks against machine learning models. In: 2017 IEEE Symposium on Security and Privacy (SP), pp. 3–18. IEEE (2017)
19. Sivaranjini, S., Sujatha, C.: Deep learning based diagnosis of Parkinson's disease using convolutional neural network. Multimed. Tools Appl. **79**(21), 15467–15479 (2020)
20. Thigpen, D., Kappler, A., Brem, R.: The role of ultrasound in screening dense breasts–a review of the literature and practical solutions for implementation. Diagnostics **8**(1), 20 (2018)
21. Wheeler, S.B., Rocque, G., Basch, E.: Benefits of breast cancer screening and treatment on mortality. JAMA **331**(3), 199–200 (2024)
22. Yi, S., Chen, Z., Yi, L., She, F.: CAS: Breast cancer diagnosis framework based on lesion region recognition in ultrasound images. J. King Saud Univ.-Comput. Inf. Sci. **35**(8), 101707 (2023)
23. Zhang, T., et al.: Radiomics and artificial intelligence in breast imaging: a survey. Artif. Intell. Rev. **56**(Suppl 1), 857–892 (2023)
24. Zhuang, Z., Yang, Z., Raj, A.N.J., Wei, C., Jin, P., Zhuang, S.: Breast ultrasound tumor image classification using image decomposition and fusion based on adaptive multi-model spatial feature fusion. Comput. Methods Programs Biomed. **208**, 106221 (2021)

Towards Robust Breast Segmentation: Leveraging Depth Awareness and Convexity Optimization For Tackling Data Scarcity

Mohammad Hossein Zolfagharnasab[1(✉)], Tiago Gonçalves[1], Pedro Ferreira[1], Maria J. Cardoso[2], and Jaime S. Cardoso[1]

[1] INESC TEC, Faculdade de Engenharia, Universidade do Porto, Rua Dr. Roberto Frias, 4200-465 Porto, Portugal
mohammad.h.zolfagharnasab@inesctec.pt

[2] Champalimaud Foundation, Avenida Brasília, 1400-038 Lisbon, Portugal

Abstract. Breast segmentation has a critical role for objective pre and postoperative aesthetic evaluation but challenged by limited data (privacy concerns), class imbalance, and anatomical variability. As a response to the noted obstacles, we introduce an encoderdecoder framework with a Segment Anything Model (SAM) backbone, enhanced with synthetic depth maps and a multiterm loss combining weighted crossentropy, convexity, and depth alignment constraints. Evaluated on a 120patient dataset split into 70% training, 10% validation, and 20% testing, our approach achieves a balanced test dice score of 98.75%—a 4.5% improvement over prior methods—with dice of 95.5% (breast) and 89.2% (nipple). Ablations show depth injection reduces noise and focuses on anatomical regions, yielding dice gains of 0.47% (body) and 1.04% (breast). Geometric alignment increases convexity by almost 3% up to 99.86%, enhancing geometric plausibility of the nipple masks. Lastly, crossdataset evaluation on CINDERELLA samples demonstrates robust generalization, with small performance gain primarily attributable to differences in annotation styles.

Keywords: Breast Segmentation · Convexity Regularization · Depth-Aware Semantic Segmentation · SAM-Driven Encoding

1 Introduction

Breast segmentation plays a critical role in reconstructive and cosmetic surgery planning, as it supports delineating anatomical landmarks that are used as priors in many pre- and post-operative aesthetic evaluation techniques [1].

With the growing demand for consistent, objective, and time-efficient evaluation tools, the automation of aesthetic assessment has garnered increasing attention in recent years [2]. Despite this necessity, achieving an accurate segmentation in real-world clinical settings remains challenging. Privacy concerns and strict regulations surrounding sensitive data limit the creation, sharing, and accessibility of diverse datasets, which impedes the development of robust, data-driven

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
T. Zhang et al. (Eds.): Deep-Breath 2025, LNCS 16142, pp. 41–51, 2026.
https://doi.org/10.1007/978-3-032-05559-0_5

solutions [3]. Additionally, class imbalance and annotation inconsistencies—particularly in small, underrepresented regions like the nipple—make these areas prone to misclassification [4]. These difficulties are exacerbated by visual variability from post-surgical scars, skin discoloration, and irregularities (such as, pimples or blemishes), ultimately reducing the reliability of aesthetic segmentation models in practical clinical use [5].

To tackle the identified challenges, this study introduces a segmentation framework, specifically designed for breast aesthetics. Our method integrates synthetic depth values—as supplementary information signals [6]—to enhance the model's perception of breast topology as well as promoting convexity to produce anatomically plausible masks for nipple. In summary, this work makes the following contributions:

1. A thorough evaluation of competitive state-of-the-art (SoTA) pre-trained segmentation models tailored for breast aesthetic segmentation;
2. Integration of depth maps to enhance geometric understanding and improve segmentation performance under limited data conditions;
3. Introduction of a combined loss formulation incorporating convexity-aware constraints to encourage anatomically plausible predictions.

To facilitate reproducibility, the implementation code for this study is publicly available on GitHub[1].

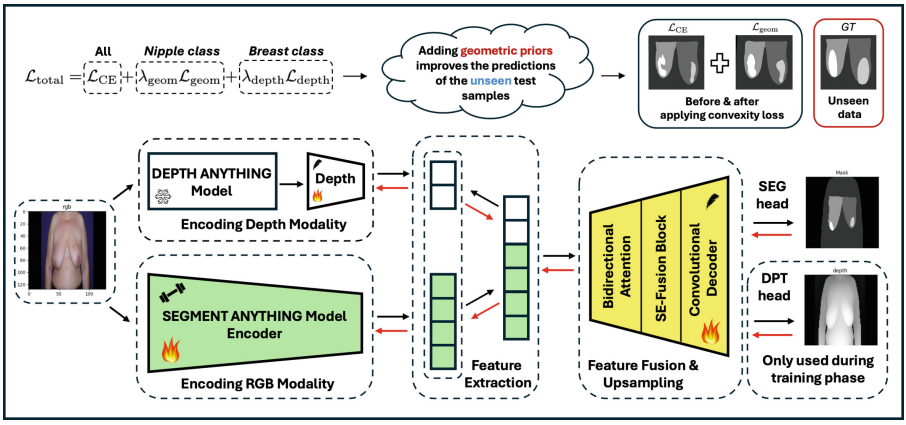


Fig. 1. Overview of the proposed model, where we combine the extracted features using SAM encoder and a lightweight depth model for improved feature representation. Next, the decoder upsamples the features using attention and depth injection layers to predict the segmentation mask as: background, body, breast, and nipple classes.

[1] https://github.com/MsainZn/DepthAwareSegmentation.

2 Related Work

To delineate the research landscape relevant to this study, we organize the related work into three thematic areas, each reflecting a component of our methodology.

SoTA in Semantic Segmentation. Semantic segmentation has seen drastic progress with the advent of deep convolutional neural networks (CNNs) and, more recently, vision transformer (ViT) architectures as models such as SegFormer [7], and Segment Anything Model (SAM) [8] set competitive benchmarks across various medical and non-medical datasets. While SoTA segmentation models perform well on general tasks, they often struggle to generalize to specialized domains due to limited/no exposure to such data during training [9].

Geometry-Aware Loss Functions. To enhance anatomical plausibility and enforce structural consistency, various geometry-aware loss functions have been proposed. Convexity constraints [10], boundary-aware losses [11], and graph-based techniques [12] have been integrated into segmentation frameworks to further promote the existing geometrical domain knowledge into the training phase.

Incorporating Depth in Segmentation. Although depth maps are not easily accessible, they offer valuable cues about surface topology that can enhance 2D segmentation, especially when color or texture cues are weak [13]. In medical imaging, depth has been used to improve segmentation in anatomies with complex geometry, such as regularizing knee arthroscopy masks [14], or segmenting skin lesions with photometric stereo cues [15]. Considering progress in fused RGB-depth models—like the work by Jamal et al. [16] on surgical scene understanding—this approach is likely to gain traction in the following years.

3 Proposed Method

Depicted in Fig. 1, our model adopts a standard encoder-decoder architecture, where the encoder extracts features from different modalities, and the decoder reconstructs the output masks using this information.

For the encoding stage, we employed SAM encoder due to its strong feature representation capabilities, with three key modifications introduced to better suit our task. First, we replaced the standard interactive decoder as it relied on additional annotations. This was unsuitable for our automatic pipeline and the bounding boxes were found suboptimal on showcasing fine-grained areas like nipple. Second, we avoided using depth as an additional encoder input, as it invalidated much of the pre-trained weights, conflicting with our goal of maximizing use of prior knowledge. Third, instead of directly passing the embeddings from depth model, we found that a small trainable encoder afterwards improves utilization of depth signals. To validate the effectiveness of our model, we compared the proposed method against prior work on the same dataset, as well as over ten SoTA pretrained models.

Concerning the decoder, our design goal was to make it lightweight and flexible. At its core, there is a multimodal bidirectional cross-attention module with learnable positional encoding, where the RGB features attend to depth and vice

versa which aims for maximizing the RBG-depth alignment. Next, the outputs are passed through a *squeeze-and-excitation* fusion block followed by a *spatial attention gate*, both designed to enhance salient features while suppressing noisy activations. Finally, the fused output passes through a lightweight convolutional decoder with two heads—for segmentation and for depth estimation. We must highlight that the depth head is only used during training to provide auxiliary supervision for which we used *Depth-Anything-v2* [13] to produce synthetic depth maps. While not metrically accurate, the synthetic depth still offers valuable geometric cues that help separate breasts from the background and body.

4 Loss Formulation

In addition to leveraging the SAM's rich encoder, it is equally important to design effective loss formulation to ensure better model generalization, notably for the nipples where we have severe class imbalance and inconsistent annotations. To do so, we propose a multi-term loss with trainable weights ($\lambda_{\text{conv}}, \lambda_{\text{dpt}} = \lambda_{\text{gdpt}}$), as in Eq. 1:

$$\mathcal{L}_{\text{total}} = -\sum_{i,j} w_{y_{i,j}} \log\left(\frac{\exp(\mathbf{S}_{i,j,y_{i,j}})}{\sum_c \exp(\mathbf{S}_{i,j,c})}\right) + \lambda_{\text{hf}} \sum_{c\in\mathcal{C}} \sum_{b=1}^{B} \frac{1}{A_{b,c}} \left\| \hat{M}_{b,c} - \mathcal{C}_{\text{attn}}(\hat{M}_{b,c}) \right\|_1 + \lambda_{\text{dpt}} \left\| D_{\text{pred}} - D_{\text{gt}} \right\|_1 + \lambda_{\text{gdpt}} \sum_{k\in\{x,y\}} \left\| \nabla_k D_{\text{pred}} - \nabla_k D_{\text{gt}} \right\|_1 . \quad (1)$$

Briefly, the *weighted cross-entropy* (WCE) loss is the term responsible for addressing the class imbalance by weighting each pixel's contribution $w_{y_{i,j}}$ based on the inverse frequency of its true class $y_{i,j}$ in the data distribution, and this weight is then applied to the predicted logits $\mathbf{S}_{i,j,c}$ for each pixel (i, j).

The second term, *convexity*, loss encourages smoother and more uniform segmentation masks for selected classes, i.e., the nipples, by applying a learned multi-scale attention-based closing gap filter $\mathcal{C}_{\text{attn}}$— using six different kernel sizes—to the predicted soft masks $\hat{M}_{b,c}$ for batch b and class C. The loss normalizes by the soft mask area $A_{b,c}$ to avoid bias towards larger regions.

Lastly, *depth losses* aligns the predicted depth maps D_{pred} with the ground truth (GT) D_{gt} using an L_1 norm for both the depth values and their spatial gradients ∇, computed via fixed Sobel filters. This encourages further edge-aware geometric consistency to enhance boundary detection between the breasts and the background. We must also report that using same weight parameter for depth and depth gradient showcases slightly better results in our work.

5 Dataset

In this study, we used *Breast-Aesthetics* (BA) [17] which comprised of 120 RGB image and segmentation masks of the front upper torso of patients. Since different models require their specific input properties, we applied tailored preprocessing functions to ensure optimal utilization of the pretrained weights [18].

This included employing fixed aspect-ratio resizing combined with black padding to preserve spatial consistency. We also computed the pixel distribution across segmentation classes—44.50% background, 42.16% body, 12.35% breasts, and 1.00% nipples—, which we later used to calculate class imbalance weights.

Next, the dataset was partitioned into train, validation, and test sets using a 70:10:20 split. Models were trained for 100 epochs with a batch size of 4, a learning rate of 10^{-5}, and the Adam optimizer— taking around 3 h on an A100 GPU. Segmentation performance was then assessed using Intersection over Union (IoU), Dice Similarity Coefficient (DSC), and convexity (Conv) [19].

6 Results and Discussion

Validation with Previous Work. To validate the effectiveness of the proposed approach, we first compared our results with prior efforts on this dataset, as summarized in Table 1. Briefly, the proposed method achieves a test balanced DSC score of 98.75%, representing an average improvement of 4.5% over previous approaches, which reported scores around 94%. This highlights a perceptible performance gain and demonstrates the effectiveness of our approach.

Table 1. Comparison with previous works on test balanced DSC.

Method	Year	DSC Score (%)	Loss formulation
Fernandes et al. [20]	2018	93.93	Ordinal
Rebelo et al. [21]	2019	94.17	Quality-based regularizer
Cruz et al. [22]	2025	94.30	Ordinal and contact surface
Ours	2025	**98.75**	Depth-aware and convexity

Model Comparison. Here we conducted a performance evaluation over various pretrained models to establish the effectiveness of our approach compared with SoTA (results in Table 2). Following similar approach as in [23], we selected challenging samples featuring high asymmetry (row-1), bruise marks (row-2), scars (row-3), and varied camera distance with slight color change (row-4) to test robustness of the system. Figure 2 highlights the advantage of pretrained weights over a U-Net trained from scratch, producing noisy masks due to limited training data. MobileViT-v2, DPT, and BEIT improved on U-Net but struggled with nipples and breast regions, due to their small dataset representation and high variability. MobileNet-v2 showed fewer artifacts than MobileViT-v2 and BEIT but misclassified bruises as nipples (row-2) with more pixelated masks (row-3,4). Segformer was the second most robust model, handling bruises, scars, and asymmetry (rows 1–3), yet sensitive to mild skin tone changes and camera distance (row-4). In the meantime, while earlier approaches produced noisy predictions, our model consistently delivers consistent segmentation masks. Achieving an almost 90% DSC on the challenging nipple class—only 1% of all pixels—highlights how the integration of SAM's rich encoder, depth maps, and geometric alignment enables precise boundary delineation and improved generalization.

Table 2. Comparison of the obtained DSC score (%) against leading SoTA models.

Region	UNet	MobileNet	MobileViT	Segformer	BEIT	DPT	UperNet	Ours
Back	99.5	99.3	98.5	99.8	99.7	98.7	99.8	**99.8**
Body	89.9	96.8	95.4	98.5	97.8	93.9	98.5	**98.7**
Breast	69.0	89.8	85.6	94.3	91.7	80.9	92.9	**95.5**
Nipple	62.7	78.1	57.6	88.8	79.2	61.5	87.3	**89.2**
Mean	91.6	96.9	95.3	98.4	97.7	94.3	98.3	**98.7**

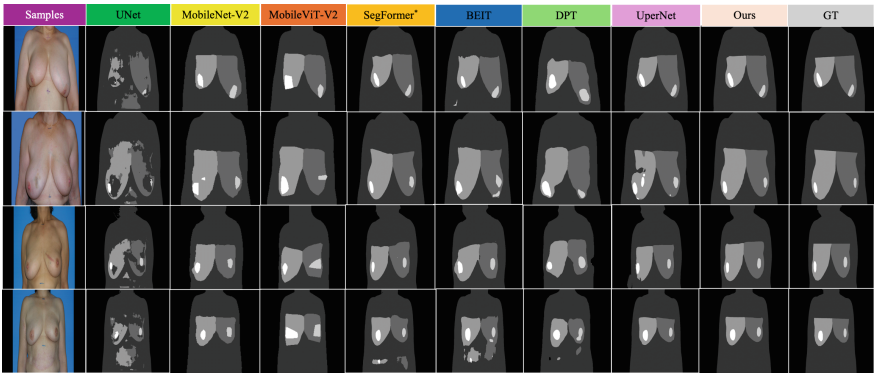


Fig. 2. A visual comparison between the performance of different fine-tined models compared to ours on *unforeseen test samples*.

Impact of Incorporating Depth. The objective of supplementing depth information is to better separate breasts from the body and background. In this regard, the simplest approach that is treating depth as an extra input channel for pretrained models proved ineffective, as it either degraded overall accuracy—an DSC decay of 6% for Segformer and 9% for MobileNet-v2— since pretrained weights had to be discarded or it was restricted by the encoder's design of some models (e.g., SAM and DPT). To overcome these drawbacks, we adopted a second strategy, where we injected depth signals via bidirectional attention immediately following the SAM encoder, and introduced a dedicated depth-supervision loss during training. As shown in Fig. 3, although quantitative gains are modest—likely due to limited test samples—depth cues help the model focus on anatomically relevant regions and improve robustness to noise.

Geometric Alignment. Concluded in model comparison, both data scarcity and class imbalance negatively impacts the predicted masks for nipple area. To address these issues, we incorporated convexity alignment into the loss formulations to encourage plausible geometric properties such as smoothness, and convexity for the predicted masks. Despite a slight decrease in metrics correlating with CE—which is expected as convexity would not necessarily promote accuracy—, Table 3 showcases a perceptible improvement in Conv metric. This can also be highlighted from, Fig. 4, where adding convexity loss improves geo-

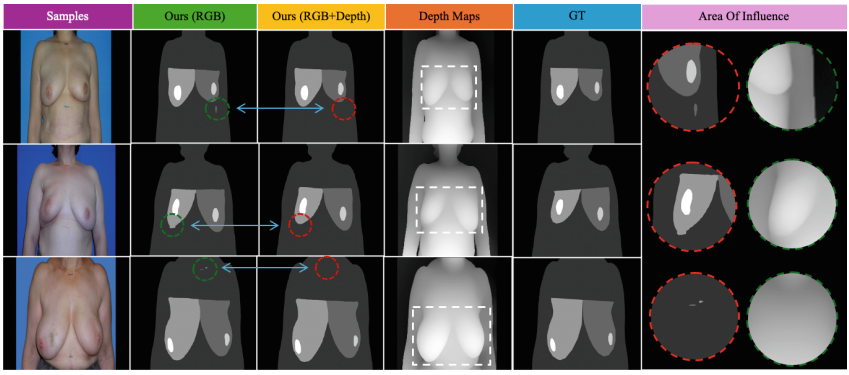


Fig. 3. A brief comparison demonstrating how incorporating depth information aids in denoising and reduces false-positive artifacts on *unforeseen test samples*. This provides an DSC improvement of 0.47% on body, and 1.04% on breast class, which may be small for our dataset but probably more perceptible in larger datasets.

metric consistency on unseen test samples— even with imbalanced training data. On the downside, it sometimes reduces mask area undesirably—though in some cases, one could argue that the predictions appear more accurate than the GT. For this study, we mitigated this side effect by incorporating area-preserving constraints, adaptive weighting, and a multi-kernel strategy into the convexity loss to better balance geometric fidelity and accuracy; however, this also highlight a promising direction for future work.

Table 3. Impact of geometric alignment on our model.

Loss	Back		Body		Breast		Nipple			Mean		
	IoU	DSC	IoU	DSC	IoU	DSC	IoU	DSC	Conv	mIoU	mDSC	Loss
CE	99.2	99.85	97.6	98.91	91.0	95.40	77.8	90.37	**97.72**	91.48	99.95	0.2412
CE+Conv	98.3	99.79	97.3	98.74	90.4	95.48	76.7	89.15	**99.86**	90.93	98.75	3.4224

Cross-Dataset Evaluation. As our final experiment, we evaluated cross-dataset generalization using CINDERELLA project [24] data, with GT masks created for 36 images (2/3 train, 1/3 test). As shown in Table 4, our model trained on *BA* demonstrated strong performance even before fine-tuning on CINDERELLA, achieving a modest improvement (7% on average) after fine-tuning, highlighting the model's strong generalizability. In the nipple region, fine-tuning yielded less than a 6% DSC gain, improving from 82% to 88%, whereas the baseline SAM achieved a DSC of 11.1%. For the breast class, which was initially expected to be problematic, further analysis—illustrated in Fig. 5—showed that differences mainly stemmed from annotation styles: CINDERELLA masks

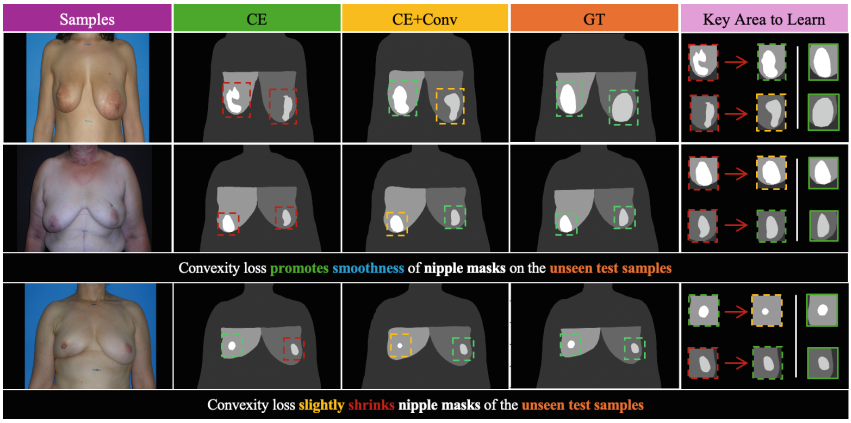


Fig. 4. Impact of geometric alignment on the accuracy predicted nipple masks.

included the armpit area, unlike the original dataset where the breast region was generally masked far below. As a result, our model was mostly learning to adapt to the extended mask boundaries rather than correcting structural inaccuracies. Considering that our training did not include data augmentation, the results here highlight the promising generalization of the proposed model.

Table 4. DSC scores on the CINDERELLA dataset.

Phase	Back	Body	Breast	Nipple	Mean
Ours (SAM-Freezed)	73.1	3.5	30.5	11.1	40.1
Ours (Fine-tuned on BA)	98.4	87.7	77.1	82.4	90.1
Ours (Fine-tuned on CINDERELLA)	99.5	97.7	95.1	88.8	97.9

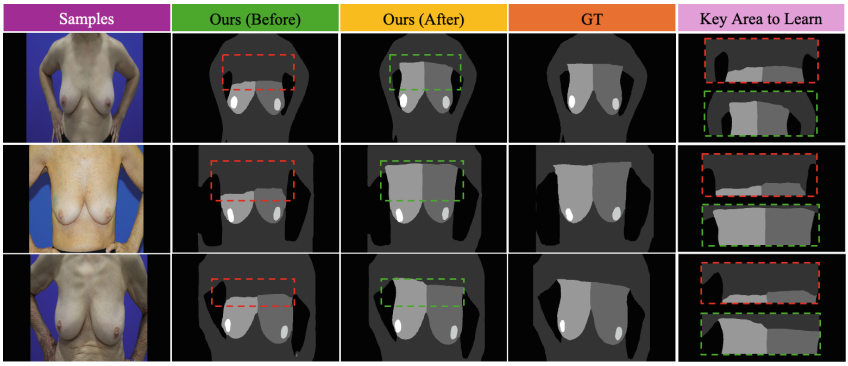


Fig. 5. Cross-dataset evaluation showcasing the model mostly learns the differences between annotations after fine-tuning on CINDERELLA dataset.

7 Conclusions and Future Work

The proposed combination of a SAMbased encoder with synthetic depth signals and simple geometric penalties improves the consistency and accuracy of breast aesthetic masks—achieving a balanced DSC of 98.75% and enhanced delineation of small structures such as nipples. Incorporating depth information helps reduce spurious activations, while convexity and ellipse losses encourage smoother and more geometrically plausible contours. Future work could expand in two key directions: first, addressing the potential negative impact of geometric losses that may overly reduce mask size, by developing adaptive or relaxed constraints that preserve critical anatomical details; and second, leveraging data augmentation and semisupervised learning techniques to utilize a larger pool of unlabeled or weakly labeled images, thereby improving model robustness and generalization.

Acknowledgments.. This work has received funding from the European Union's Horizon Europe research and innovation programme under the Grant Agreement 101057389-CINDERELLA project, and by the Portuguese Foundation for Science and Technology (FCT) through the Ph.D. Grant "2024.06378.BD".

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

Data Policy. The dataset used for this study is private, but it can be accessed upon request.

References

1. Zolfagharnasab, M.H., et al.: Predicting aesthetic outcomes in breast cancer surgery: a multimodal retrieval approach. In: Mann, R.M., et al. (eds.) Deep-Breath 2024. LNCS, vol. 15451, pp. 137–147. Springer, Cham (2025). https://doi.org/10.1007/978-3-031-77789-9_14
2. Freitas, N., Montenegro, H., Cardoso, M.J., Cardoso, J.S.: Reproducing asymmetries caused by breast cancer treatment in pre-operative breast images. In: 2024 IEEE International Symposium on Biomedical Imaging (ISBI), pp. 1–5. IEEE (2024)
3. Cardoso, J.S., Silva, W., Cardoso, M.J.: Evolution, current challenges, and future possibilities in the objective assessment of aesthetic outcome of breast cancer locoregional treatment. Breast **49**, 123–130 (2020)
4. Silva, W., Carvalho, M., Mavioso, C., Cardoso, M.J., Cardoso, J.S.: Deep aesthetic assessment and retrieval of breast cancer treatment outcomes. In: Pinho, A.J., Georgieva, P., Teixeira, L.F., Sánchez, J.A. (eds.) IbPRIA 2022. LNCS, vol. 13256, pp. 108–118. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-04881-4_9
5. Montenegro, H., Cardoso, M.J., Cardoso, J.S.: A disentangled approach to predict the aesthetic outcomes of breast cancer treatment. In: Del Bue, A., Canton, C., Pont-Tuset, J., Tommasi, T. (eds.) ECCV 2024. LNCS, vol. 15631, pp. 311–327. Springer, Cham (2025). https://doi.org/10.1007/978-3-031-91838-4_19

6. Zolfagharnasab, M.H., Bahrani, M., Saghayan, M.H., Masoumi, F.S.: Exploring a novel multi-channel structure to improve facial expression recognition on occluded samples using deep convolutional neural network. J. Artif. Intell. Appl. Innov. **1**(2), 26–41 (2024)
7. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: SegFormer: simple and efficient design for semantic segmentation with transformers. Adv. Neural Inf. Process. Syst. **34**, 12077–12090 (2021)
8. Kirillov, A., et al.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4015–4026 (2023)
9. Freitas, N., Silva, D., Mavioso, C., Cardoso, M.J., Cardoso, J.S.: Deep edge detection methods for the automatic calculation of the breast contour. Bioengineering **10**(4), 401 (2023)
10. Van Kaick, O., Fish, N., Kleiman, Y., Asafi, S., Cohen-Or, D.: Shape segmentation by approximate convexity analysis. ACM Trans. Graph. (TOG) **34**(1), 1–11 (2014)
11. Borse, S., Wang, Y, Zhang, Y., Porikli, F.: InverseForm: a loss function for structured boundary-aware segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5901–5911 (2021)
12. Royer, L.A., Richmond, D.L., Rother, C., Andres, B., Kainmueller, D.: Convexity shape constraints for image segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 402–410 (2016)
13. Yang, L., et al.: Depth anything V2. Adv. Neural. Inf. Process. Syst. **37**, 21875–21911 (2024)
14. Liu, F., Jonmohamadi, Y., Maicas, G., Pandey, A.K., Carneiro, G.: Self-supervised depth estimation to regularise semantic segmentation in knee arthroscopy. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12261, pp. 594–603. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59710-8_58
15. Goyal, M., Oakley, A., Bansal, P., Dancey, D., Yap, M.H.: Skin lesion segmentation in dermoscopic images with ensemble deep learning methods. IEEE Access **8**, 4171–4181 (2019)
16. Jamal, M.A., Mohareri, O.: Rethinking RGB-D fusion for semantic segmentation in surgical datasets (2024)
17. Cardoso, J.S., Cardoso, M.J.: Towards an intelligent medical system for the aesthetic evaluation of breast cancer conservative treatment. Artif. Intell. Med. **40**(2), 115–126 (2007)
18. Zolfagharnasab, M.H., Damari, S.: A comparative analysis of machine learning models in news categorization. U. Porto J. Eng. **10**(3), 23–38 (2024)
19. Akiva, P., Dana, K., Oudemans, P., Mars, M.: Finding berries: segmentation and counting of cranberries using point supervision and shape priors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 50–51 (2020)
20. Fernandes, K., Cardoso, J.S.: Ordinal image segmentation using deep neural networks. In: 2018 International Joint Conference on Neural Networks (IJCNN), pp. 1–7. IEEE (2018)
21. Rebelo, J., Fernandes, K., Cardoso, J.S.: Quality-based regularization for iterative deep image segmentation. In: 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 6734–6737. IEEE (2019)
22. Cruz, R.P.M., Cristino, R., Cardoso, J.S.: Learning ordinality in semantic segmentation. IEEE Access (2025)

23. Zolfagharnasab, M.H., PourMohammadBagher, L., Bahrani, M.: Intelligent travel recommendations using neural collaborative filtering for touristic landmarks of Iran. J. Data Sci. Model. 119–147 (2025)
24. Kaidar-Person, O., et al.: Evaluating the ability of an artificial-intelligence cloud-based platform designed to provide information prior to locoregional therapy for breast cancer in improving patient's satisfaction with therapy: the CINDERELLA trial. PLOS ONE **18**(8), 1–12 (2023)

Divergence-Aware Training with Automatic Subgroup Mitigation for Breast Tumor Segmentation

Eleonora Poeta[1(✉)], Luisa Vargas[2], Daniele Falcetta[2], Vincenzo Marciano'[2], Eliana Pastor[1], Tania Cerquitelli[1], Elena Baralis[1], and Maria A. Zuluaga[2]

[1] Politecnico di Torino, Turin, Italy
{Eleonora.Poeta,Eliana.Pastor,Tania.Cerquitelli,Elena.Baralis}@polito.it
[2] EURECOM, Biot, France
{Luisa.Vargas,Daniele.Falcetta,Vincenzo.Marciano,
Maria.Zuluaga}@eurecom.fr

Abstract. Deep learning models for breast tumor segmentation in DCE-MRI may exhibit disparities in performance across demographic and clinical subgroups, raising concerns about fairness and clinical trustworthiness. In this work, we propose a subgroup-aware in-processing mitigation strategy that integrates divergence-based regularization directly into the training loop. By leveraging interpretable metadata (e.g., age, menopausal status, breast density), we identify subgroups where the model underperforms and assign higher loss weights to these samples in proportion to their divergence from average performance. Our method enables the model to focus training on underrepresented or harder-to-segment subpopulations without requiring external data or post-processing correction. We evaluate our approach on the MAMA-MIA 2025 challenge dataset, demonstrating improvements in both overall segmentation quality and fairness score. Our results highlight the potential of in-processing mitigation as an effective and practical pathway toward equitable medical image segmentation.

Keywords: Fairness · Breast Tumor Segmentation · Bias Mitigation · Subgroup Analysis

1 Introduction

Breast tumor delineation using dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI) plays a crucial role in the diagnosis, treatment planning, and monitoring of breast cancer [15,32]. Advances in deep learning techniques for medical image segmentation have led to state-of-the-art models that achieve remarkable accuracy in the task [22]. However, these models may be susceptible to algorithmic unfairness, meaning they perform unevenly across patient subgroups based on protected attributes such as ethnicity, age, or socioeconomic

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
T. Zhang et al. (Eds.): Deep-Breath 2025, LNCS 16142, pp. 52–62, 2026.
https://doi.org/10.1007/978-3-032-05559-0_6

background [8,23,28]. These disparities undermine the model's fairness and the overall trust in AI systems.

In-processing strategies have been proposed to improve fairness dynamically during model training, without relying on external data or post-hoc adjustments [26,31]. Existing techniques typically focus on mitigating disparities related to a single sensitive attribute, such as sex [2,19], age, or clinical site [27]. However, focusing solely on one attribute at a time may overlook disparities arising at the intersection of multiple factors. Moreover, these methods often require a priori knowledge of which subgroups to monitor or protect.

In this work, we propose FairMedSeg, an in-processing mitigation strategy tailored for medical image segmentation. FairMedSeg automatically identifies subgroups that exhibit *divergent* performance, defined as statistically significant deviations from the population-level average [17], and reduces such disparities during training. We achieve this by extending the divergence-based reweighting framework proposed initially by Koudounas et al. [10] for speech processing, and adapting it to the specific challenges of 3D medical image segmentation. We validate our method on the MAMA-MIA 2025 challenge dataset [5], which provides annotated DCE-MRI breast scans enriched with clinical metadata, demonstrating that FairMedSeg improves segmentation accuracy while reducing fairness gaps across clinically meaningful subgroups.

Our approach introduces several key innovations for in-processing fairness mitigation in this domain: (i) an automatic and intersectional subgroup discovery pipeline tailored to clinical metadata; (ii) a fully dynamic, performance-driven sample reweighting mechanism integrated directly into the training loop; (iii) a test-time-agnostic design that eliminates the need for metadata at inference, since fairness is addressed during training; (iv) the removal of any dependence on predefined sensitive attributes, enabling fairness-aware learning without prior subgroup specification.

2 Related Work

Fairness-enhancing strategies can be broadly categorized into *pre-processing*, *post-processing*, and *in-processing* methods [4,12].

Pre-processing techniques modify the training data to reduce biases before model learning. Strategies include dataset rebalancing [11], generative data augmentation [18], and harmonization techniques to eliminate confounding variables [21]. Causal perturbation methods create counterfactual samples by altering sensitive attributes while maintaining clinical validity [20]. Although model-agnostic and easy to implement, these methods may distort natural data distributions or struggle to generalize across domains.

Post-processing methods operate on model outputs or internal predictions after training is complete, without altering the model architecture or input data. Common techniques include subgroup-specific threshold adjustment [6] and post-hoc confidence calibration [14]. While effective in some cases, these methods

typically rely on access to metadata at inference time, an assumption that may not hold in real-world clinical deployments.

In-processing strategies incorporate fairness constraints into the training procedure. Adversarial approaches [31] learn representations that are predictive of the target task while invariant to sensitive attributes, but can be unstable, compromising task performance, particularly in high-dimensional data. Group Distributionally Robust Optimization [25] and fairness-aware regularization [26] penalize disparities across predefined subgroups, yet require accurate group labels and can struggle to scale when subgroup definitions are noisy or evolving. Moreover, they are primarily developed for classification tasks and may not translate directly to structured prediction settings such as medical image segmentation.

Divergence-based regularization [10] is a recent in-processing mitigation technique that dynamically adjusts training based on disparities among subgroups. It identifies subgroups whose performance deviates from the overall model accuracy and assigns higher weights to the samples from those groups. This encourages the model to focus on learning from the divergent subgroups. So far, it has been used in the context of speech processing. In this work, we adapt the divergence-aware framework to 3D medical image segmentation, extending the method to automatically discover vulnerable subgroups based on clinical metadata. Unlike classical reweighting approaches based on group frequency statistics [9] or meta-learned strategies [29], our method leverages subgroup-level performance divergence as a dynamic signal to guide training. In contrast to focal-loss-based techniques [1,30], which prioritize sample-level difficulty, our approach explicitly addresses performance disparities across clinically meaningful subgroups defined by metadata.

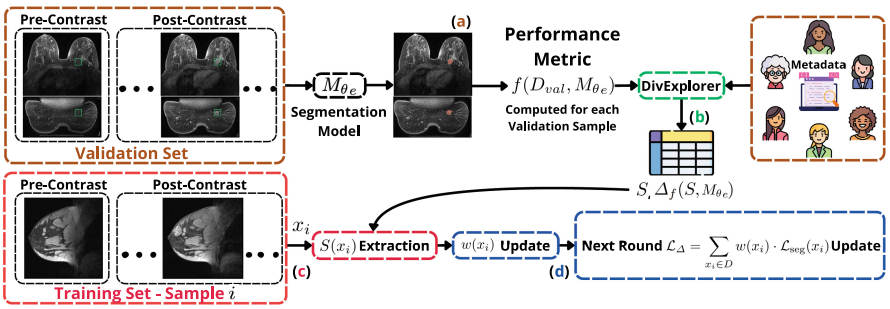


Fig. 1. Overview of FairMedSeg weight update scheme. After each training epoch e, performance is evaluated on a validation set (a), and metadata-defined subgroups are discovered using `DivExplorer` (b). Sample weights are updated based on subgroup divergence, prioritizing underperforming subgroups (c). These weights are used to compute a fairness-aware loss for the next epoch, iteratively reducing subgroup disparities throughout training (d).

3 Method

3.1 Problem Formulation

Let $D = \{(x_i, y_i, m_i)\}_{i=1}^{N}$ denote a dataset of N annotated DCE-MRIs, where x_i is the input image, y_i the corresponding ground truth segmentation mask, and m_i a vector of clinical metadata attributes (e.g., age, menopausal status, breast density), i.e., $m_i = (m_{i,1}, \ldots, m_{i,k})$, with k the number of attributes. A segmentation model M_θ is trained to predict $\hat{y}_i = M_\theta(x_i)$ by minimizing a segmentation loss function $\mathcal{L}_{\text{seg}}$ over the training data.

We define a *subgroup* S as a subset of samples sharing specific metadata characteristics, i.e., $S = \{(x_i, y_i, m_i) \in D \,|\, m_{i,j_1} = m^*_{j_1} \wedge m_{i,j_2} = m^*_{j_2} \wedge \ldots \wedge m_{i,j_p} = m^*_{j_p}\}$ with $j_1, j_2, ..., j_p$ the indices of the features within m_i. For instance, a subgroup can be defined by $\{$*age = 41–50*, *menopausal status = pre*$\}$. Let $f(S, M_\theta)$ represent the model's performance on subgroup S (e.g., mean Dice score), and $f(D, M_\theta)$ its performance over the full dataset. The divergence of subgroup S is [16]:

$$\Delta_f(S, M_\theta) = f(S, M_\theta) - f(D, M_\theta), \tag{1}$$

In our setting, where we measure model performance using the average Dice score, negative values indicate performance below the global average. Our goal is to reduce divergence while preserving overall segmentation accuracy by defining a training strategy that is sensitive to subgroup disparities and can proactively mitigate them.

3.2 FairMedSeg Mitigation Strategy

Figure 1 provides an overview of the training process for FairMedSeg. During each training epoch e, the resulting model M_{θ_e} is evaluated using a validation set, which allows us to assess the model's performance. By exploiting the metadata available at training (e.g., age, menopausal status, breast density), we automatically identify subgroups and estimate their divergence. The estimated performance gaps for each subgroup are then used to update sample weights and adjust the loss function in the subsequent epoch. By reducing fairness gaps at training, the final model M_θ is test-time-agnostic by eliminating the need for metadata at inference.

Subgroup Discovery. We define subgroups as conjunctions of interpretable clinical metadata attributes (e.g., menopausal status, age group, breast density). To systematically construct these subgroups, we leverage the `DivExplorer` framework [16], which enumerates all combinations of metadata attributes that satisfy a minimum support threshold, denoted as *minsup*. This threshold ensures that each subgroup is sufficiently represented in the dataset, enabling reliable estimation of performance metrics.

Divergence-Based Sample Weighting. We assign training sample weights based on subgroup-level performance divergences identified during step (b). Each

training sample x_i may belong to one or more metadata-defined subgroups, denoted as $S(x_i) = \{S \in \mathcal{S} \mid x_i \in S\}$. For example, a 45-year-old patient in pre-menopause belongs not only to the marginal subgroups $\{age = 45\}$ and $\{menopausal\ status = pre\}$ but also to the intersectional subgroup $\{age = 45, menopausal\ status = pre\}$. After each training epoch, we evaluate the model of current epoch e (M_{θ_e}) on the validation set and compute the divergence score $\Delta_f(S, M_{\theta_e})$ for each subgroup $S \in \mathcal{S}$ (Eq. 1).

To update training priorities, we assign each training sample a weight equal to the maximum absolute divergence across the subgroups to which it belongs:

$$w(x_i) = \max_{S \in S(x_i)} |\Delta_f(S, M_{\theta_e})|. \tag{2}$$

The updated weights are then used in the next training epoch via $\mathcal{L}_{\text{total}}$ (Eq. 4). This weighting mechanism increases the influence of samples from subgroups where the model exhibits the most divergent behavior, encouraging the model to focus on those more challenging or underperforming samples.

Loss Formulation. The loss function is composed of two terms: segmentation loss $\mathcal{L}_{\text{seg}}$ and divergence-aware loss $\mathcal{L}_{\Delta}$. Specifically, they are defined as:

$$\mathcal{L}_{\Delta} = \sum_{x_i \in D} w(x_i) \cdot \mathcal{L}_{\text{seg}}(x_i), \tag{3}$$

where $\mathcal{L}_{\text{seg}}(x_i)$ denotes the standard segmentation loss (e.g., Dice Cross-entropy, Dice Focal) for the training sample x_i. The final loss function combines the original segmentation loss with the divergence-aware component:

$$\mathcal{L}_{\text{total}} = \alpha \cdot \mathcal{L}_{\text{seg}} + (1 - \alpha) \cdot \mathcal{L}_{\Delta}, \tag{4}$$

where $\alpha \in [0, 1]$ balances the trade-off between overall segmentation performance and subgroup fairness mitigation.

4 Experimental Setup

We evaluate our method using three widely used backbone 3D medical image segmentation architectures implemented within the MONAI framework [3]. These are: UNet [24], a baseline convolutional encoder–decoder architecture; SegResNet [13], a residual convolutional network optimized for volumetric medical data; and SwinUNETR [7], a hybrid transformer-based model that combines a Swin Transformer encoder with a UNet-like decoder. For each backbone architecture considered, we compare the base model (trained with a standard segmentation loss) to its fairness-enhanced counterpart using FairMedSeg.

4.1 Dataset and Data Preparation

We evaluate our method on the MAMA-MIA dataset [5], which comprises 1,506 T1-weighted DCE-MRI cases from female patients diagnosed with breast cancer,

divided into training and validation sets. Each case comprises one pre-contrast and up to five post-contrast dynamic phases, acquired across multiple clinical centers under varying imaging protocols. Tumor regions were manually segmented by a panel of 16 expert radiologists, ensuring high-quality ground truth annotations of the primary lesions. In addition to imaging data, the dataset includes rich clinical metadata. We focus on three key attributes, patient age, menopausal status, and breast density, to define clinically meaningful subgroups for fairness evaluation, as specified by the challenge guidelines.

We preprocess each DCE-MRI case by first reorienting the images to RAS (Right-Anterior-Superior) and resampling them to an isotropic resolution of 1.0 mm using bilinear interpolation for the images and nearest-neighbor interpolation for the labels. We concatenate the pre-contrast and two post-contrast phases into a single 3-channel 3D volume. To ensure consistent spatial dimensions across the dataset, we standardize all inputs to a fixed shape of $320 \times 320 \times 128$ by applying cropping or padding as needed. To focus learning on the relevant anatomy, we apply a foreground cropping step based on non-zero voxel intensity. We apply a set of data augmentation strategies during training. Namely, random flipping along each spatial axis (x, y, z) with a probability of 0.5, intensity normalization restricted to non-zero voxels, and random intensity scaling and shifting with factors and offsets up to $\pm 10\%$.

For post-processing the model output, we refine each predicted segmentation by leveraging patient-specific anatomical priors. Specifically, we use metadata-provided bounding box coordinates that delineate the breast region for each case. These coordinates are used to apply a spatial mask that restricts predictions to the anatomically plausible area, effectively removing any segmented regions outside the breast tissue.

4.2 Evaluation Metrics

We follow the evaluation protocol from the MAMA-MIA challenge, relying on five metrics: the Dice Similarity Coefficient (DSC), the Normalized 95th Percentile Hausdorff Distance (NormHD), the Performance Score (PS), the Fairness Score (FS), and the Total Score (TS).

The DSC quantifies the spatial overlap between predicted and ground truth segmentation masks. The NormHD assesses boundary-level accuracy by computing the 95th percentile Hausdorff Distance, normalized by the image resolution and clipped to the range $[0, 1]$. The PS summarizes segmentation quality as the average of the DSC and the complement of the NormHD:

$$\mathrm{PS} = 0.5 \cdot (\mathrm{DSC} + (1 - \mathrm{NormHD})) \,.$$

The FS assesses fairness across variables. For each variable $v \in \mathcal{V}$, the challenge computes a disparity measure D_v based on subgroup-level performance variations, defining FS as:

$$\mathrm{FS} = \frac{1}{|\mathcal{V}|} \sum_{v=1}^{|\mathcal{V}|} (1 - D_v),$$

with $|\mathcal{V}| = 3$, the three clinically relevant variables: *age*, *menopausal status*, and *breast density*. Finally, PS and FS are aggregated using equal weighting into the Total Score, i.e., $\text{TS} = (1 - \alpha) \cdot \text{PS} + \alpha \cdot \text{FS}$, with $\alpha = 0.5$.

5 Experimental Results

Quantitative Results. Table 1 summarizes the results using the best-performing configuration for each model using the validation set within MAMA-MIA. We also report the result of the best model on the official MAMA-MIA private test set, as evaluated by the challenge organizers. In particular, we found experimentally that a regularization weight of $\alpha = 0.2$ and a $minsup = 0.001$ value for `DivExplorer` provided the most consistent improvements across architectures.

On the validation set, we observe substantial gains in TS (+14.0). On the official test set (Table 1, highlighted row) the UNet with FairMedSeg achieves a FS of 0.86. Overall, these results demonstrate that FairMedSeg achieves strong generalization across different patient groups while ensuring fairness is preserved or enhanced. These results confirm that fairness-aware training can reduce disparities among subgroups without compromising overall accuracy.

Table 1. FairMedSeg-enhanced and base models' performance on MAMA-MIA's validation and test sets. The highlighted row refers to the model submitted to the *MAMA-MIA* challenge (private test set). **Bold** denotes the best scores on the validation set.

Model	Method	DSC	NormHD	Perf. Score	Fairness Score	Total Score
UNet	Base (val)	0.7251	0.1380	0.7935	0.8334	0.8135
	FairMedSeg (val)	0.7691	0.0960	0.8364	**0.8580**	0.8472
	FairMedSeg (test)	0.66	0.16	0.75	0.86	0.80
SegResNet	Base (val)	0.7453	0.1226	0.8113	0.7414	0.7763
	FairMedSeg (val)	**0.7865**	**0.0806**	**0.8529**	0.8445	**0.8475**
SwinUNETR	Base (val)	0.6421	0.1970	0.7225	0.6480	0.6853
	FairMedSeg (val)	0.7478	0.1236	0.8121	0.8379	0.8250

Subgroup Analysis with Heatmaps. To better understand subgroup-specific disparities, we visualize the average DSC across combinations of clinical metadata (Fig. 2) for the UNet model using FairMedSeg evaluated on the MAMA-MIA test set. The left plot shows DSC stratified by age group and breast density. The right one shows DSC by age group and menopausal status.

We observe substantial variability in segmentation performance across intersectional subgroups. For instance, patients aged 41–50 with breast density category `a`, exhibit the lowest DSC (0.158), while older patients (71+) generally achieve higher DSCs, especially in high-density categories. Similarly, the

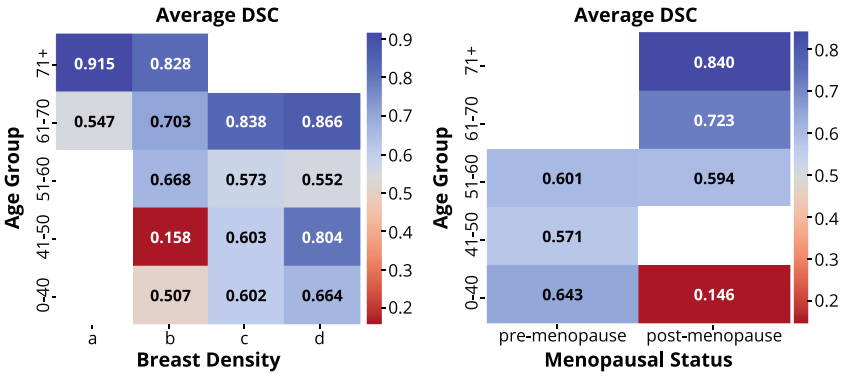


Fig. 2. Subgroup-Level DSC Heatmaps. *Left:* Average DSC by age group and breast density (a–d). *Right:* Average DSC by age group and menopausal status.

right heatmap highlights substantial differences in the case of post-menopausal patients aged 0–40, which achieves only 0.146 DSC, compared to 0.643 for their premenopausal counterparts. These findings confirm that subgroup disparities are not only present but also intersectional in nature, motivating the need for intersectionality-aware training strategies such as FairMedSeg.

6 Discussion and Conclusions

We introduced FairMedSeg, a novel in-processing fairness mitigation strategy tailored for medical image segmentation. Our results demonstrate that FairMedSeg can significantly reduce subgroup disparities in breast tumor segmentation without compromising overall accuracy. By leveraging metadata-defined subgroups and dynamically adjusting training priorities based on divergence, our method effectively directs learning toward underperforming cohorts.

Automatic Subgroup Identification. A key strength of FairMedSeg lies in its ability to identify and mitigate fairness gaps without requiring prior subgroup definitions or manual intervention. This is particularly advantageous in clinical scenarios where vulnerable populations may be unknown or evolve over time. Moreover, the divergence-aware regularization can be flexibly integrated into a wide range of segmentation backbones, as demonstrated in our evaluation across convolutional and transformer-based architectures.

Intersectional Subgroup Disparities. The subgroup-level heatmap analysis (Fig. 2) reveals that performance disparities are not uniformly distributed but often emerge at specific intersections of clinical metadata. For example, patients aged 41–50 with breast density `a` and postmenopausal patients under age 40, achieve the lowest DSC values. These gaps are often difficult to detect or mitigate using fairness strategies that treat attributes independently. FairMedSeg's capacity to autonomously discover and prioritize such subgroups underscores its

practical value in real-world medical AI systems. While Fig. 2 illustrates persistent disparities in certain intersectional subgroups, our method reduces these gaps compared to the baseline. However, achieving full equity across all combinations remains a challenge due to limited sample sizes in rare subgroups.

Limitations and Future Work. We have identified some pending limitations. First, the subgroup divergence computation depends on validation performance, which may be unstable for low-sample subgroups. Moreover, our current sample weighting relies on the maximum subgroup divergence per sample, which may oversimplify intersectional subgroup dynamics. Incorporating richer aggregation mechanisms represents a promising direction for future refinement. Second, we focused on three metadata variables (age, menopausal status, and breast density) due to the MAMA-MIA challenge rules. Extending our approach to additional metadata, e.g., ethnicity, hormonal receptor status, or treatment history, is a promising future direction. Finally, although we fixed the regularization weight ($\alpha = 0.2$), automatic tuning could further enhance generalization across datasets.

Acknowledgments. This work is partly supported by the ANR-BMBF TRAIN (ANR-22-FAI1-0003-02) and by the FAIR - Future Artificial Intelligence Research (Piano Nazionale Di Ripresa e Resilienza (PNRR) – Missione 4 Componente 2, Investimento 1.3 – D.D. 1555 11/10/2022, PE00000013) and the spoke "FutureHPC & BigData" of the ICSC - Centro Nazionale di Ricerca in High-Performance Computing, Big Data and Quantum Computing, both funded by the European Union - NextGenerationEU.

Disclosure of Interests. This manuscript reflects only the authors' views and opinions; neither the European Union nor the European Commission can be considered responsible for them. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Abraham, N., Khan, N.M.: Adaptive focal loss for sparse object detection. arXiv preprint arXiv:1810.07842 (2021)
2. Alsulaimawi, A.: Enforcing fairness in deep medical imaging via adversarial learning. In: 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), pp. 2557–2560. IEEE (2021)
3. Cardoso, M.J., et al.: MONAI: an open-source framework for deep learning in healthcare. arXiv preprint arXiv:2211.02701 (2022)
4. Chen, R.J., et al.: Algorithmic fairness in artificial intelligence for medicine and healthcare. Nat. Biomed. Eng. **7**(6), 719–742 (2023)
5. Garrucho, L., et al.: A large-scale multicenter breast cancer DCE-MRI benchmark dataset with expert segmentations. Sci. Data **12**(1), 453 (2025)
6. Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. In: Advances in Neural Information Processing Systems, vol. 29 (2016)
7. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D.: Swin UNEtr: Swin transformers for semantic segmentation of brain tumors in MRI images. In: International MICCAI Brainlesion Workshop, pp. 272–284. Springer (2021)

8. Huti, M., Lee, T., Sawyer, E., King, A.P.: An investigation into race bias in random forest models based on breast DCE-MRI derived radiomics features. In: Workshop on Clinical Image-Based Procedures, pp. 225–234. Springer (2023)
9. Kamiran, F., Calders, T.: Data preprocessing techniques for classification without discrimination. Knowl. Inf. Syst. **33**(1), 1–33 (2012)
10. Koudounas, A., Pastor, E., de Alfaro, L., Baralis, E.: Mitigating subgroup disparities in speech models: a divergence-aware dual strategy. IEEE Trans. Audio Speech Lang. Process. (2025)
11. Larrazabal, A.J., Nieto, N., Peterson, V., Milone, D.H., Ferrante, E.: Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. Proc. Natl. Acad. Sci. **117**(23), 12592–12594 (2020)
12. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. ACM Comput. Surv. (CSUR) **54**(6), 1–35 (2021)
13. Myronenko, A.: 3D MRI brain tumor segmentation using autoencoder regularization. In: International MICCAI Brainlesion Workshop, pp. 311–320. Springer (2018)
14. Noriega-Campero, A., Bakker, M.A., Garcia-Bulle, B., Pentland, A.: Active fairness in algorithmic decision making. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, pp. 77–83 (2019)
15. Park, G.E., Kim, S.H., Nam, Y., Kang, J., Park, M., Kang, B.J.: 3D breast cancer segmentation in DCE-MRI using deep learning with weak annotation. J. Magn. Reson. Imaging **59**(6), 2252–2262 (2024)
16. Pastor, E., De Alfaro, L., Baralis, E.: Looking for trouble: analyzing classifier behavior via pattern divergence. In: Proceedings of the 2021 International Conference on Management of Data, pp. 1400–1412 (2021)
17. Pastor, E., Gavgavian, A., Baralis, E., Alfaro, L.: How divergent is your data? Proc. VLDB Endowment **14**(12), 2835–2838 (2021)
18. Paxton, K., Aslansefat, K., Thakker, D., Papadopoulos, Y.: Evaluating fairness and mitigating bias in machine learning: a novel technique using tensor data and Bayesian regression. arXiv preprint arXiv:2506.11627 (2025)
19. Pérez-García, F., Sparks, R., Ouyang, J., et al.: Fairness in deep cardiac MRI segmentation: assessing sex and racial bias in deep learning-based segmentation. Med. Image Anal. **80**, 102358 (2022)
20. Pfohl, S.R., Duan, T., Ding, D.Y., Shah, N.H.: Counterfactual reasoning for fair clinical risk prediction. In: Machine Learning for Healthcare Conference, pp. 325–358. PMLR (2019)
21. Puyol-Antón, E., et al.: Fairness in cardiac MR image analysis: an investigation of bias due to data imbalance in deep learning based segmentation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, 27 September–1 October 2021, Proceedings, Part III 24, pp. 413–423. Springer (2021)
22. Rayed, M.E., Islam, S.S., Niha, S.I., Jim, J.R., Kabir, M.M., Mridha, M.: Deep learning for medical image segmentation: state-of-the-art advancements and challenges. Inf. Med. Unlocked **47**, 101504 (2024)
23. Ricci Lara, M.A., Echeveste, R., Ferrante, E.: Addressing fairness in artificial intelligence for medical imaging. Nat. Commun. **13**(1), 4581 (2022)
24. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 234–241. Springer (2015)

25. Sagawa, S., Koh, P.W., Hashimoto, T.B., Liang, P.: Distributionally robust neural networks. In: International Conference on Learning Representations (2020). https://openreview.net/forum?id=ryxGuJrFvS
26. Sarhan, M.H., Navab, N., Eslami, A., Albarqouni, S.: Fairness by learning orthogonal disentangled representations. In: Computer Vision–ECCV 2020: 16th European Conference, Part XXIX, vol. 16, pp. 746–761. Springer (2020)
27. Seyyed-Kalantari, L., Zhang, H., McDermott, M., Chen, I., Ghassemi, M.: CheXclusion: fairness gaps in deep chest X-ray classifiers. Sci. Rep. **11**(1), 10452 (2021)
28. Xu, Z., Li, J., Yao, Q., Li, H., Zhao, M., Zhou, S.K.: Addressing fairness issues in deep learning-based medical image analysis: a systematic review. npj Digit. Med. **7**(1), 286 (2024)
29. Yan, H., Zou, J., et al.: FORML: fairness-optimized robust meta-learning. arXiv preprint arXiv:2202.01719 (2022)
30. Yeung, M., et al.: Unified focal loss: generalising dice and cross entropy-based losses to handle class imbalanced medical image segmentation. arXiv preprint arXiv:2102.04525 (2022)
31. Zhao, B., Xiao, X., Gan, G., Zhang, B., Xia, S.T.: Maintaining discrimination and fairness in class incremental learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13208–13217 (2020)
32. Zhou, L., et al.: Prototype learning guided hybrid network for breast tumor segmentation in DCE-MRI. IEEE Trans. Med. Imaging (2024)

Towards Early Detection: AI-Based Five-Year Forecasting of Breast Cancer Risk Using Digital Breast Tomosynthesis Imaging

Manon A. Dorster[1,2], Felix J. Dorfner[1,3], Mason C. Cleveland[1], Melisa S. Guelen[1,3], Jay Patel[1], Dania Daye[1,4], Jean-Philippe Thiran[2], Albert E. Kim[1(✉)], and Christopher P. Bridge[1]

[1] Athinoula A. Martinos Center for Biomedical Imaging, Mass General Brigham and Harvard Medical School, Charlestown, MA, USA
{akim46,cbridge}@mgh.harvard.edu
[2] École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland
[3] Charité - Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Berlin, Germany
[4] Department of Radiology, University of Wisconsin School of Medicine and Public Health, Madison, WI, USA

Abstract. As early detection of breast cancer strongly favors successful therapeutic outcomes, there is major commercial interest in optimizing breast cancer screening. However, current risk prediction models achieve modest performance and do not incorporate digital breast tomosynthesis (DBT) imaging, which was FDA-approved for breast cancer screening in 2011. To address this unmet need, we present a deep learning (DL)-based framework capable of forecasting an individual patient's 5-year breast cancer risk directly from screening DBT. Using an unparalleled dataset of 161,753 DBT examinations from 50,590 patients, we trained a risk predictor based on features extracted using the Meta AI DINOv2 image encoder, combined with a cumulative hazard layer, to assess a patient's likelihood of developing breast cancer over five years. On a held-out test set, our best-performing model achieved an AUROC of 0.80 on predictions within 5 years. These findings reveal the high potential of DBT-based DL approaches to complement traditional risk assessment tools, and serve as a promising basis for additional investigation to validate and enhance our work.

1 Introduction

Breast cancer (BC) is the leading cause of cancer-related mortality in women worldwide [11]. Given recent advances in targeted and endocrine therapies, early diagnosis of breast cancer is strongly linked to higher curative potential in women affected by the disease. Consequently, there is a major interest in developing

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
T. Zhang et al. (Eds.): Deep-Breath 2025, LNCS 16142, pp. 63–71, 2026.
https://doi.org/10.1007/978-3-032-05559-0_7

optimal screening tools that can both maximize early detection and reduce false-positive assessments. Current risk assessment models (e.g., Tyrer-Cuzick model) [13] evaluate known risk factors, such as family history and breast density, but achieve modest accuracy. To this end, computational models employing 2D mammography, which contain orthogonal data compared to conventional risk factors, have demonstrated superior performance over traditional risk assessment models in predicting future breast cancer risk [16]. Therefore, there is considerable interest in optimizing end-to-end DL models, with breast imaging data given as input, to better forecast future breast cancer risk.

Digital breast tomosynthesis (DBT) was FDA approved for breast cancer screening in 2011, and offers advantages over 2D mammography. 2D mammography is limited in its ability to reveal small tumors due to superimposed breast tissues that may mask abnormalities, while DBT imaging offers near-3D imaging capabilities [9]. Current work in the field of AI prediction models based on breast imaging mainly applies to 2D mammography data [14,16]. Further studies leverage mammography imaging data from multiple time points, thus exploiting longitudinal variability of image features [4,7,8,15]. A study led by Eriksson et al. [5] explores one-year breast cancer risk prediction based on DBT input data. A five-year time frame is recommended for breast cancer risk prediction [1]; however, no long-term risk prediction model based on DBT data currently exists. Therefore, using an unparalleled dataset of DBTs from Mass General Brigham, we present a DL-based prototype of a 5-year risk assessment model, leveraging the pre-trained Meta AI DINOv2 [10] image feature encoder followed by a custom fully-connected layer.

2 Methods

2.1 Dataset

For model training, development, and testing, we compiled a dataset of 161,753 screening DBT examinations from 50,590 patients at Mass General Brigham. All DBTs included in the dataset are **screening** DBTs conducted on female patients between January 2012 and November 2023. Moreover, the dataset is divided into two cohorts: the **pre-cancer** group, consisting of 6,449 DBT examinations from 2,915 patients prior to biopsy-confirmed breast cancer, and the **healthy** group, which comprises of 155,304 DBT examinations from 47,675 patients unaffected by breast cancer. As most patients received multiple DBT examinations, we defined a data point in our dataset as a **DBT study** conducted for a given patient at a given time point. The **study date** or time $\mathbf{t_0}$ of a DBT study is the time at which the examination was conducted. For the pre-cancer cohort, the **diagnosis date** is the date at which the patient was first diagnosed with biopsy-confirmed breast cancer. Year 1 spans from the study date (t_0) to 365 days later, year 2 from $t_1 = t_0$+365 days to 365 days after t_1, etc.

A positive case is defined as a DBT study conducted within five years prior to a breast cancer diagnosis. Diagnosis dates in the pre-cancer cohort were extracted

using the Qwen 2.5 72B large language model [17] on all available breast pathology reports. Our implementation excludes DBT studies conducted within less than 6 months of breast cancer diagnosis from the validation and test sets. For the healthy cohort, since some DBT studies are recent and lack full 5-year follow-up, we retained only negative cases with at least one year of screening follow-up. Combined with the label masking process (cf. Section 2.1), this ensures that the healthy dataset includes only confirmed negatives, excluding DBT studies from patients lost to follow-up or who may develop cancer in the future.

Only studies with available right and left cranio-caudal and medio-lateral oblique views (RCC, LCC, RMLO, and LMLO) were included. Each view acquired during a DBT study is saved as an individual DICOM file, which we refer to as a **DBT series** in the rest of this work. The dataset was divided into training, validation, and test subsets, corresponding to 70%, 15% and 15% of the data, respectively. All studies for a unique patient were assigned to the same subset. Due to the high imbalance between the pre-cancer and healthy groups, which reflects natural breast cancer occurrence among women, we used a weighted sampler in order to load the data from the two groups in equal proportions during training.

Labeling and Masking. The output of the risk prediction model is a 5-element vector, each element corresponding to the predicted **breast cancer status** for years 1 through 5. Labeling and masking are performed using a **sliding window** approach, similar to the framework in [16], such that each DBT study for a given patient receives its own label and mask corresponding to the five-year period after acquisition of the study. Hence, the risk assessment model is robust to variable amounts of follow-up. A label at year n indicates whether the patient had a positive (1) or negative (0) cancer status at that specific time point, as shown in Eq. 1. If breast cancer is diagnosed at year n, the label is also marked positive for years $n+1$ to 5.

$$\mathbf{y} = [y_1, ..., y_5], \quad \text{with } y_i = \begin{cases} 0, & \text{if negative at year } i, \\ 1, & \text{if positive at year } i, \end{cases} \quad \text{for } i \in \{1, ..., 5\} \tag{1}$$

Variability in the amount of follow-up among patients in the dataset calls for a masking strategy. We defined a mask vector $\mathbf{w}$ for each study with five binary entries, each indicating whether breast cancer status is available for the corresponding year. Years with no information regarding the malignancy status of a given patient were masked out during training, validation, and inference.

Dataset Statistics. Patients in the dataset display age ranges of 68.8 ± 11.7 years and 65.1 ± 10.9 years for the pre-cancer and healthy cohorts, respectively. All DBT studies were acquired using Hologic Selenia Dimensions systems. The demographics of the dataset are detailed in Table 1.

Breast radiology reports assign scores using the Breast Imaging Reporting and Data System (BI-RADS) [18], which also classifies breast density into four

Table 1. Demographics of the full dataset. Percentages are relative to each column subset. For "years to cancer", percentages are within pre-cancer cases only.

Breast Density (Patient Counts)	**Healthy**		**Pre-cancer**
a - almost entirely fatty	4,785 (10%)		117 (4%)
b - scattered areas of fibroglandular density	22,929 (48%)		1,424 (49%)
c - heterogeneously dense	17,371 (37%)		1,060 (36%)
d - extremely dense	2,537 (5%)		104 (4%)
Unknown	53 (<1%)		210 (7%)
Ethnicity (Patient Counts)	**Healthy**		**Pre-cancer**
White	38,458 (81%)		2,467 (85%)
Black	2,762 (6%)		129 (4%)
Asian	2,156 (4%)		117 (4%)
Other/Unknown	4,299 (9%)		202 (7%)
BC Subtypes (Patient Counts)	**Train**	**Validation**	**Test**
Invasive ductal carcinoma (IDC)	1,691 (69%)	158 (69%)	166 (71%)
Invasive lobular carcinoma (ILC)	289 (12%)	30 (13%)	28 (12%)
IDC & ILC	137 (5%)	11 (5%)	7 (3%)
Ductal carcinoma in situ (DCIS)	186 (8%)	16 (7%)	17 (7%)
Other	148 (6%)	14 (6%)	17 (7%)
Years to Cancer (Study Counts)	**Train**	**Validation**	**Test**
1 year	1,718 (31%)	18 (4%)	24 (5%)
2 years	1,182 (22%)	132 (28%)	154 (31%)
3 years	999 (18%)	126 (27%)	121 (24%)
4 years	813 (15%)	103 (22%)	104 (21%)
5 years	774 (14%)	86 (19%)	95 (19%)
BI-RADS (Study Counts)	**Training**	**Validation**	**Test**
0 - need for additional imaging	7,790 (6%)	1,206 (6%)	1,157 (5%)
1 - negative	93,495 (80%)	17,646 (80%)	17,477 (81%)
2 - benign	16,899 (14%)	3,092 (14%)	2,958 (14%)
3 - probably benign	15 (<1%)	2 (<1%)	7 (<1%)
≥4 - suspicious, suggestive of malignancy	9 (<1%)	0 (0%)	0 (0%)

categories [12] (cf. Table 1). High density has been shown to correlate with increased breast cancer risk [2]. A few density labels in our dataset are missing (cf. Table 1), thus we excluded those cases from the density subgroup analysis (Table 3). Screening cases are rarely attributed a BI-RADS score above 3 (see Table 1); suspicious findings are typically labeled BI-RADS 0 and are referred for diagnostic imaging or biopsy.

2.2 Risk Assessment Model

Image feature representations were obtained using the pre-trained DINOv2 foundation model (see Fig. 1). Using this state-of-the-art image encoder allowed for rapid model prototyping and reduced the substantial computational resources needed for development of a task-specific model. Embeddings were saved to disk and used for experiments in which we fed the feature representations to a linear layer, termed the **cumulative hazard layer**, which outputs one risk score per year following the study date.

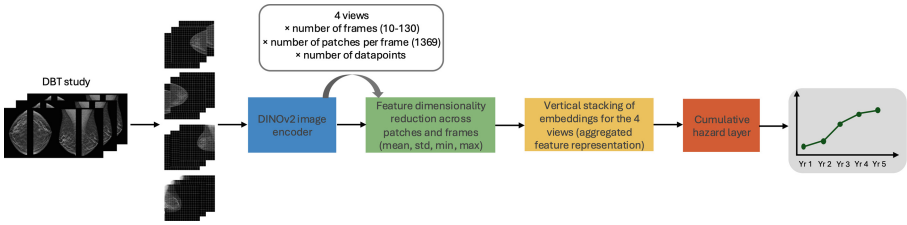


Fig. 1. 5-year risk assessment model pipeline, inspired from Fig. 1 in [16].

Image Encoding and Feature Extraction. Each DBT study is composed of four views, with each view (DBT series) containing multiple 2D slices (**frames**). For every frame of a DBT series, we resized each frame to 518 by 518 pixels and ran a forward pass through the DINOv2 ViT-B/14 image encoder, from which we omitted the classifier head. Each forward pass yields one CLS token and 1369 patch token embeddings per frame, all of size 768. Embeddings for all frames in a DBT series are then stacked to obtain a feature representation of the full DBT series. Given that there are four DBT series per DBT study (RCC, LCC, RMLO, and LMLO), this results in a substantial number of features per study. To reduce dimensionality while attempting to preserve informative content, we computed summary statistics–mean, standard deviation (SD), minimum, and maximum–of each feature across patches within each frame and across frames within each DBT series when using the patch token representations. Following standard vision transformer architectures, one CLS embedding, typically used for downstream image-level classification tasks, is extracted for each 2D frame and captures global characteristics of the image. By contrast, patch embeddings capture patch-level representations. For each DBT series, we computed the summary statistics across frames. Different combinations of patch or CLS summary statistic embeddings were used for the downstream training of the cumulative hazard layer, as displayed in Table 2.

Cumulative Hazard Layer. Similarly to the implementation in [16], the cumulative hazard layer is a fully connected linear layer that takes as input the aggregated representation of a DBT study, consisting of the stacked feature representations of all four DBT series in the study. The cumulative hazard layer then outputs five predicted risk probabilities for each of the corresponding years following the study date.

We define ground truth labels $\mathbf{y}$ as described in Eq. 1. The risk probabilities over 5 years should follow a monotonically increasing trend. It would not be coherent for a predicted risk at year n to be higher than that at year $n+1$, hence the use of a cumulative risk probability framework. A fully-connected linear layer is applied to the feature representation of a DBT study and yields five output values. These logit outputs are passed through the softplus activation function to ensure non-negative values, which are then cumulatively summed across the years, producing monotonically increasing risk scores. Since the softplus function

outputs values in the range $[0, \infty)$, a final normalization step is applied to map the cumulative predictions into valid probability values within the range $[0, 1]$.

Given an aggregated feature representation of a DBT study x, the probability that the concerned patient will develop breast cancer k years after the study date is given by the sum of the independent risk probabilities for years 1 through k. Equation 2 summarizes the cumulative risk probability layer output at a given year k, $\hat{y}_i$ being the model risk prediction at year i:

$$P(t_{cancer} = k|x) = p_k = 1 - \exp(-\sum_{i=1}^{k} \hat{y}_i), \quad 1 \leq k \leq 5 \tag{2}$$

In the model training experiments, we aimed to minimize the risk prediction objective defined as the sum of the binary cross-entropy (BCE) losses for each year weighted by the mask $\mathbf{w}$, averaged over N samples in a batch (Eq. 3).

$$\mathcal{L}_{BCE} = \frac{1}{N} \sum_{i=1}^{N} \frac{\sum_{j=1}^{5} w_{i,j} \left[y_{i,j} \log p_{i,j} + (1 - y_{i,j}) \log(1 - p_{i,j}) \right]}{\sum_{j=1}^{5} w_{i,j}} \tag{3}$$

3 Results and Discussion

AUROC results for the different risk assessment models we developed are shown in Table 2. We observed that models based on CLS embeddings, which learn a global representation of an image [3], displayed overfitting trends, with the validation loss measured during model training starting to increase after the first few epochs. Therefore, we find that using patch tokens allows for the model to better capture detailed features [3] relating to local regions that are relevant to risk prediction (see Table 2). Furthermore, better performance is achieved when considering the entire distribution over patch embeddings by calculating mean, standard deviation, minimum and maximum statistics over all patches (**model 3** defined in Table 2) rather than solely the mean (model 1 defined in Table 2). AUROC values based on density subgroups computed with model 3 predictions on test set data (Table 3) highlight the robustness of the model across clinically relevant subgroups.

The Kaplan-Meier curves in Fig. 2 address whether cases predicted as high-risk at year n develop cancer earlier than those predicted as low-risk at year n or across all 5 years. These trends demonstrate a clear separation between cancer-free survival in each group and highlight the model's ability to efficiently stratify the DBT studies into risk groups.

To our knowledge, our dataset of 50,590 patients and 161,753 screening DBTs is one of the largest used to train a breast cancer risk prediction model using end-to-end DL. Our work demonstrates promise in applying DL to screening DBTs to quantify 5-year future risk. Interestingly, we note that the performance of our model is comparable to state-of-the-art risk models [16] employing 2D mammograms as input. Therefore, given the emerging availability of DBTs for

Table 2. Yearly AUROC metrics for risk assessment model predictions on the test set (498 positive, 21,101 negative DBT studies). Cell colors represent the best (green), second best (orange), and worst (red) AUROC for each year. *Metrics considered unreliable due to small sample (24 positives).

Model Name	Embed. Type	Summary Stats	Year 1*	Year 2	Year 3	Year 4	Year 5
Model 1	patch	mean	0.6682	0.7082	0.7495	0.7775	0.7941
Model 2	patch	mean, SD	0.6663	0.7063	0.7514	0.7776	0.7915
Model 3	patch	mean, SD, min, max	0.6758	0.7175	0.7544	0.7886	0.7999
Model 4	CLS	mean	0.6746	0.6993	0.7399	0.7779	0.7908
Model 5	CLS	mean, SD	0.6980	0.7087	0.7363	0.7680	0.7856
Model 6	CLS	mean, SD, min, max	0.6923	0.6934	0.7291	0.7658	0.7837

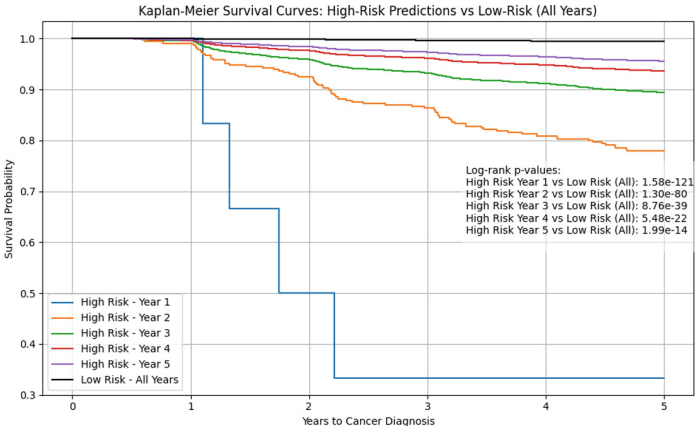


Fig. 2. Kaplan-Meier survival curves showing the estimated cancer-free survival probabilities over 5 years for different risk groups based on model 3 predictions on test set data. "Low Risk": breast cancer status predicted negative for all 5 years; "High Risk - Year n": breast cancer status predicted positive at year n. A binarization threshold for all years (0.3027) was computed by maximizing Youden's J-statistic.

breast cancer screening, further efforts are warranted to optimize DBT-based models for risk prediction.

Leveraging the existing DINOv2 model allowed for rapid model development on this large dataset and achieved promising proof-of-concept results, but training a task-specific network (either a convolutional neural network (CNN) or vision transformer) may allow for better performance. Our efforts to train such a network were limited by the available computational resources. Future work will focus on comparisons to trained, task-specific image encoders tailored for DBT imaging to evaluate the suitability of DINOv2 as an image encoder for our task. Moreover, as DINOv2 was pre-trained on natural 2D images and lacks 3D spatial awareness, it is not ideally suited to capture the pseudo-3D character-

Table 3. Breast density subgroup analysis based on model 3 predictions on test set data. Density labels: a - almost entirely fatty, b - scattered areas of fibroglandular density, c - heterogeneously dense, d - extremely dense. *Small number of positive cases, to interpret with caution.

Breast Density	**a***	**b**	**c**	**d***
# Patients Total	529 (10%)	2,657 (49%)	1,913 (35%)	301 (6%)
# Pre-cancer Patients	6 (3%)	110 (51%)	82 (38%)	16 (8%)
# Healthy Patients	523 (10%)	2,547 (49%)	1,831 (35%)	285 (6%)
# DBT Studies	1,996	10,951	7,470	1,119
Year 1 AUROC	0.8417	0.4587	0.7177	0.8083
Year 2 AUROC	0.8853	0.7259	0.7200	0.6521
Year 3 AUROC	0.8486	0.7624	0.7513	0.6564
Year 4 AUROC	0.8373	0.8037	0.7731	0.6963
Year 5 AUROC	0.8668	0.8174	0.7762	0.7166

istics of DBT volumes. To address this limitation, we are currently exploring self-supervised learning methods leveraging large amounts of unlabeled DBT images to develop pre-training strategies capable of capturing the essence of DBT data.

Another promising avenue would be more sophisticated feature dimensionality reduction methods that better characterize biologically relevant regions of DBT volumes (e.g., attention-based multiple instance learning for feature selection) [6], rather than simply using the summary statistics of the patch or CLS embeddings. In addition, as multi-modal models have demonstrated improved performances compared to image-only models [16], we also aim to pursue our work by integrating both DBT data and clinical risk factors in our model, such as age and hormonal factors, similarly to the Mirai framework in [16]. Finally, we note that our dataset is comprised predominantly of White breast cancer patients (see Table 1). To this end, future studies should recruit a more ethnically diverse dataset in order to augment model generalizability and robustness.

4 Conclusion

Using an unparalleled dataset of 50,590 patients and 161,753 screening DBTs, our study presents one of the first efforts to develop a DL-based model capable of reliably predicting five-year breast cancer risk using DBT imaging. Our study demonstrates that DBT, a modality routinely used in clinical screening, contains predictive imaging biomarkers of future cancer development that could possibly outperform those derived from 2D mammograms. The model achieves promising performances across all five years, highlighting the potential of DBT-based risk models to support personalized screening strategies. Future work will focus on improving model performance through DBT-specific pre-training and incorporating multi-modal clinical information. Ultimately, this work represents

a significant step toward the integration of AI-driven risk prediction into routine breast cancer screening, paving the way for optimized use of clinical resources aimed at reducing late-stage diagnoses and improving survival outcomes.

References

1. Bevers, T.B., et al.: Breast cancer risk reduction, version 2.2015. J. Natl. Compr. Canc. Netw. (2015)
2. Brown, A.L., Vijapura, C., Patel, M., Cruz, A.D.L., Wahab, R.: Breast cancer in dense breasts: detection challenges and supplemental screening opportunities. RSNA (2023). https://doi.org/10.1148/rg.230024
3. Caron, M., et al.: Emerging properties in self-supervised vision transformers. Comput. Vis. Pattern Recogn. (2021). https://doi.org/10.1109/ICCV48922.2021.00951
4. Dadsetan, S., Arefan, D., Berg, W.A., Zuley, M.L., Sumkin, J.H., Wu, S.: Deep learning of longitudinal mammogram examinations for breast cancer risk prediction. Pattern Recogn. (2022)
5. Eriksson, M., et al.: A risk model for digital breast tomosynthesis to predict breast cancer and guide clinical care. Sci. Transl. Med. (2022)
6. Ilse, M., Tomczak, J.M., Welling, M.: Attention-based deep multiple instance learning. In: Proceedings of the 35th International Conference on Machine Learning, vol. 80. PMLR, Stockholm, Sweden (2018)
7. Karaman, B.K., Dodelzon, K., Akar, G.B., Sabuncu, M.R.: Longitudinal mammogram risk prediction. In: MICCAI (2024)
8. Lee, H., et al.: Enhancing breast cancer risk prediction by incorporating prior images. In: MICCAI (2023)
9. McDonald, E.S., et al.: Effectiveness of digital breast tomosynthesis compared with digital mammography - outcomes analysis from 3 years of breast cancer screening. JAMA Oncol. (2016). https://doi.org/10.1001/jamaoncol.2015.5536
10. Oquab, M., et al.: Dinov2: learning robust visual features without supervision. Trans. Mach. Learn. Res. (2024). https://doi.org/10.48550/arXiv.2304.07193
11. Organization, W.H.: Global cancer burden growing, amidst mounting need for services, https://www.who.int/news/item/01-02-2024-global-cancer-burden-growing--amidst-mounting-need-for-services, Accessed: 12 Mar 2025
12. T, R., A, M., B, D.M., et al.: Breast density (classification). Reference article, Radiopaedia.orhttps://doi.org/10.53347/rID-12808, Accessed 23 Jun 2025
13. Tyrer, J., Duffy, S.D., Cuzick, J.: A breast cancer prediction model incorporating familial and personal risk factors. Stat. Med. (2004). https://doi.org/10.1002/sim.1668
14. Wang, X., et al.: Ordinal learning: Longitudinal attention alignment model for predicting time to future breast cancer events from mammograms. In: MICCAI (2024)
15. Wang, X., et al.: Predicting up to 10 year breast cancer risk using longitudinal mammographic screening history. medRxiv (2023)
16. Yala, A., et al.: Toward robust mammography-based models for breast cancer risk. Sci. Transl. Med. (2021). https://doi.org/10.1126/scitranslmed.aba4373
17. Yang, et al.: Qwen 2.5. https://github.com/QwenLM/Qwen2.5, Accessed 24 Jun 2025
18. Zonderland, H., Smithuis, R.: Bi-rads for mammography and ultrasound 2013 - updated version. https://radiologyassistant.nl/breast/bi-rads/bi-rads-for-mammography-and-ultrasound-2013, Accessed 24 Jun 2025



Large Kernel MedNeXt for Breast Tumor Segmentation and Self-normalizing Network for pCR Classification in Magnetic Resonance Images

Toufiq Musah[1,2(✉)]

[1] Department of Computer Engineering, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana
toufiqmusah32@gmail.com
[2] Global Health and Infectious Disease Group, Kumasi Centre for Collaborative Research in Tropical Medicine, Kumasi, Ghana

Abstract. Accurate breast tumor segmentation in dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI) is important for downstream tasks such as treatment planning and pathological complete response (pCR) assessment. In this work, we address both segmentation and pCR classification using the large-scale MAMA-MIA DCE-MRI dataset. We employ a large-kernel MedNeXt architecture with a two-stage training strategy that expands the receptive field from $3 \times 3 \times 3$ to $5 \times 5 \times 5$ kernels using the UpKern algorithm. This approach allows stable transfer of learned features to larger kernels, improving segmentation performance on the unseen validation set. An ensemble of large-kernel models achieved a Dice score of 0.67 and a normalized Hausdorff Distance (NormHD) of 0.24. For pCR classification, we trained a self-normalizing network (SNN) on radiomic features extracted from the predicted segmentations and first post-contrast DCE-MRI, reaching an average balanced accuracy of 57%, and up to 75% in some subgroups. Our findings highlight the benefits of combining larger receptive fields and radiomics-driven classification while motivating future work on advanced ensembling and the integration of clinical variables to further improve performance and generalization. (Code available at: https://github.com/toufiqmusah/caladan-mama-mia.git).

Keywords: Breast Tumor Segmentation · DCE-MRI · Pathological Complete Response · MedNeXt · Self-Normalizing Networks

1 Introduction

Breast cancer is a major health concern [11], and Magnetic Resonance Imaging (MRI) plays an important role in its assessment, especially for preoperative staging monitoring treatment response, and pathological complete response (pCR)

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
T. Zhang et al. (Eds.): Deep-Breath 2025, LNCS 16142, pp. 72–80, 2026.
https://doi.org/10.1007/978-3-032-05559-0_8

assessment [10,15,19]. T1-weighted dynamic contrast-enhanced MRI (DCE-MRI) can highlight tumor vascularity by using contrast agents, aiding in the localization of malignant lesions. Accurate segmentation of tumor boundaries on these images is clinically valuable, as it enables quantitative evaluation of tumor size, shape, and volume over time [2,23]. High-quality tumor segmentation further facilitates advanced analyses such radiomics feature extraction [24] for other downstream tasks [4,6], including pCR assessment [17]. There is however difficulty in obtaining expert-labeled tumor masks for breast MRI [9].

Recently, a large-scale multi-center breast DCE-MRI dataset was released to address the data scarcity. Garrucho et al. [9] introduced the MAMA-MIA challenge and dataset, the largest of its kind, comprising 1506 pre-treatment DCE-MRI cases with expert segmentations of primary breast tumors and non-mass enhancement, along with class labels for pCR status.

Deep learning architectures such as U-Net [20] and its variants [5,16,18] have been particularly successful in biomedical image segmentation. These models typically adopt conventional CNN design principles, using 3×3 convolutional kernels to balance computational efficiency and parameter count [7]. While effective, small kernels constrain the receptive field, potentially limiting the model's ability to capture long-range context, which is an important consideration in whole-breast MRI volumes [21]. To address this, recent research has explored Transformer-based architectures in medical imaging [8]. Vision Transformers and hybrid Transformer-CNN models have achieved state-of-the-art performance across various segmentation tasks by modeling long-range dependencies explicitly [14]. However, Transformers generally require large-scale training datasets to generalize well, as they lack the strong spatial inductive biases inherent to CNNs [26]. Further, the task of classifying pCR has proven difficult in both imaging and non-imaging approaches [1,22]. Radiomics [24] has proven to be capable in extracting semantic features for such downstream tasks [4,17], though feature distributions are highly abstract. Self-Normalizing Networks (SNNs) [13] were introduced to mitigate this problem by having neurons that automatically converge to zero mean and unit variance without the explicit need for normalization layers.

In this work, we propose a large-kernel MedNeXt architecture [21] for breast tumor segmentation, and a radiomics-based Self-Normalizing Network for pathological complete response (pCR) classification in DCE-MRI. Our segmentation method is motivated by the need to capture both the subtle enhancement patterns of tumors across multiple post-contrast time points and the broader breast tissue context, which larger receptive fields naturally accommodate. To effectively utilize large kernels without overfitting, we adopt a two-stage training strategy: we first train a MedNeXt model with conventional $3 \times 3 \times 3$ kernels and then expand to $5 \times 5 \times 5$ kernels using the UpKern algorithm [21]. UpKern initializes large-kernel networks by trilinear interpolation of the smaller-kernel model's convolutional filters, mitigating the performance degradation observed when large kernels are trained from scratch. For pCR prediction, we extract radiomic features from the predicted segmentations and DCE-MRI data and

train an SNN classifier to stabilize feature learning without explicit normalization layers. This combined approach targets both robust tumor segmentation and reliable downstream pCR assessment.

2 Materials and Methods

2.1 Dataset

We used the MAMA-MIA DCE-MRI dataset [9] of 1506 cases for training. Each case in the dataset includes a series of T1-weighted DCE-MRI volumes acquired at multiple time points: one pre-contrast and up to five post-contrast phases. For our experiments, we selected the pre-contrast image and the first two post-contrast images for each case, as shown in Fig. 1. Preprocessing and data handling were performed using the default nnUNet framework [12]. This included resampling all images to isotropic voxel spacing of 1.0 mm and applying z-score normalization on a per-channel basis using non-zero voxels.

The dataset comprises both unilateral and bilateral breast DCE-MRI scans. During training, we adopted nnUNet's patch-based sampling strategy with a fixed input size of $128 \times 128 \times 128$ voxels. To ensure sufficient exposure to tumor regions across both scan types, more than one-third of the patches in each batch were enforced to contain at least one foreground voxel [12]. At inference time, segmentation was performed using overlapping sliding-window patches, ensuring full-volume coverage regardless of laterality.

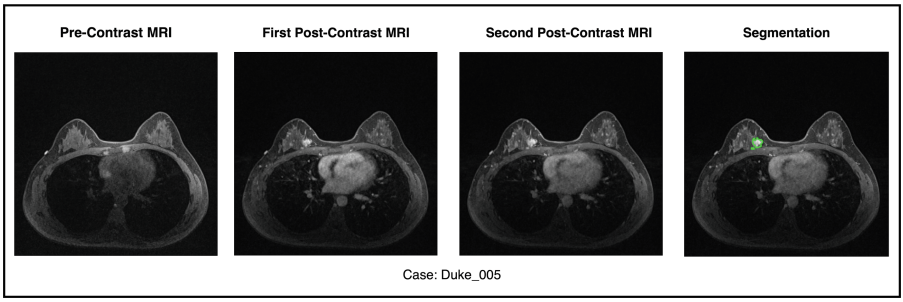


Fig. 1. Sample DCE-MRI data showing pre-contrast and post-contrast scans along with expert segmentation in green (Color figure online)

2.2 MedNeXt Segmentation Training

We employed the medium variant of MedNeXt, providing a balance between model capacity and computational efficiency for a dataset of this scale. Our modeling strategy consisted of a two-stage process. In the first stage, we trained the MedNeXt model configured with $3 \times 3 \times 3$ convolutional kernels (M^3) with

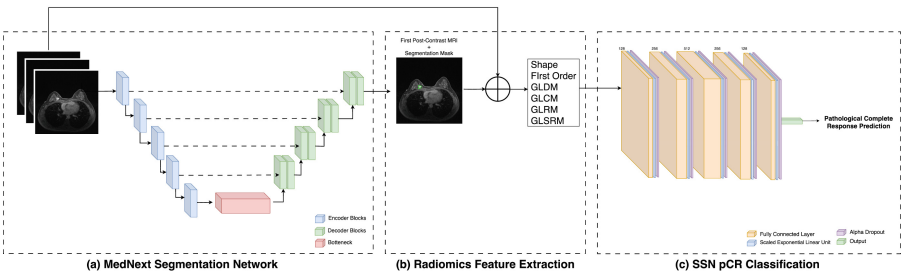


Fig. 2. Overview of the proposed pipeline. (a) MedNeXt performs tumor segmentation on DCE-MRI. (b) Radiomic features are extracted from the segmentations and first post-contrast MRI. (c) A Self-Normalizing Network (SNN) predicts pathological complete response (pCR) from the selected features.

deep supervision, stochastic gradient descent optimization, a cosine annealing learning rate, and Dice-cross-entropy loss. The training was conducted using a five-fold cross-validation split. Each fold was trained independently for 250 epochs. Following the completion of cross-validation, we identified the fold that achieved the highest mean Dice coefficient on its respective validation set, whose weights were used in the second stage of training.

In the subsequent stage, we employed the UpKern strategy to resample the learned $3 \times 3 \times 3$ kernel weights into $5 \times 5 \times 5$ kernels (${M^5}_{Base}$) via trilinear interpolation. This approach enabled a smooth transition to a large-kernel configuration, thus expanding the effective receptive field of the network without introducing instability often associated with training large kernels from random initialization. The newly initialized large-kernel MedNeXt-M model was then fine-tuned for making use of the entire training set, with all other architectural and training settings held constant.

Additionally, we generated a second large-kernel model (${M^5}_{Focal}$) by repeating the UpKern procedure and fine-tuning, this time using a combination of Dice Cross-entropy and Focal loss to increase penalization of small lesion segmentation errors.

2.3 pCR Classification

For pCR classification, we extracted radiomic features from tumor segmentations and the first post-contrast MRI (Fig. 2 (b)). To reduce redundancy, we first removed highly correlated features based on pairwise correlation analysis (threshold >90%). A sequential feature selector was then employed to identify an informative subset, resulting in 40 features spanning shape, first-order statistics, and texture descriptors.

The selected features included shape metrics such as surface area, elongation, and axis lengths; first-order statistics capturing intensity distributions, including entropy, kurtosis, and total energy; and texture-based descriptors derived from gray-level co-occurrence matrix (GLCM), gray-level size zone matrix (GLSZM),

gray-level run length matrix (GLRLM), gray-level dependence matrix (GLDM), and neighboring gray tone difference matrix (NGTDM).

We trained a Self-Normalizing Network (SNN) classifier [13] on the selected radiomic features (Fig. 2 (c)). The SNN architecture was chosen for its ability to maintain stable activations without explicit normalization layers, which is particularly advantageous when working with radiomics data that may present diverse value ranges and distributions. The model was trained using binary cross-entropy loss to predict pCR status.

3 Results

3.1 Segmentation Performance

The in-training validation results for M^3 across the 5-fold cross-validation are presented in Table 1.

Table 1. Dice scores across cross-validation folds for M^3

Fold 0	Fold 1	Fold 2	Fold 3	Fold 4	Average
0.786	0.750	0.828	0.795	0.790	0.789

Performance of M^3 was further evaluated on the unseen validation set on the Codabench platform[1]. The results are summarized in Table 2. M^3 achieved a Dice score of 0.64 and a normalized Hausdorff Distance (NormHD) of 0.3. Upon applying the UpKern strategy and fine-tuning the best performing single model, ${M^5}_{Base}$ improved the Dice score to 0.66 and the NormHD to 0.29. Further ensembling with ${M^5}_{Focal}$ led to minimal increase in Dice to 0.67, and NormHD to 0.24. The reported baseline is a 5-fold nnUNet ensemble by [9], also trained on the 1506 training dataset and evaluated on the unseen validation set.

Table 2. Performance comparison between models. Metrics are reported as Dice Score and Normalized Hausdorff Distance (95th percentile). "Parameters" denotes number of trainable parameters in a single model instance (in millions)

Model Configuration	Dice Score (↑)	NormHD (↓)	Parameters (↓) (M)
Baseline [9]	0.65	0.30	140
M^3 (5-Folds)	0.64	0.30	**30.8**
${M^5}_{Base}$	0.66	0.29	32.1
${M^5}_{Base}$ + ${M^5}_{Focal}$	**0.67**	**0.24**	✓

[1] https://www.codabench.org/competitions/7425/#/results-tab.

3.2 pCR Classification Performance

The Self-Normalizing Network (SNN) classifier achieved a balanced accuracy of 57% on the unseen validation set. The model outputs were highly confident, with most predictions clustered at 0 or 1, suggesting potential overconfidence [3]. Subgroup analysis revealed variations in performance across demographic and clinical variables. For example, balanced accuracy ranged from 75% in the 51–60 age group to 30% in the 71+ age group, and from 75% in breast density group 'a' to 41.7% in density group 'd', indicating performance disparity across subpopulations.

4 Discussion

In-training validation Dice score of M^3 averaged 0.7899 ± 0.012 across five folds, indicates stable learning and internal consistency in cross-validation. On the unseen validation set, the M^3 5-Fold ensemble achieved a Dice score of 0.64 and a NormHD of 0.30. While this represents a reasonable baseline, the performance gap relative to internal validation suggests the presence of increased data variability in the external set. Performance improved after applying the UpKern strategy to expand the receptive field. Fine-tuning the best-performing fold into ${M^5}_{Base}$ led to a Dice score of 0.66 and a reduction in NormHD to 0.29. Further ensembling ${M^5}_{Base}$ with ${M^5}_{Focal}$, which was trained using Dice-cross-entropy and Focal loss, resulted in a Dice score of 0.67 and a NormHD of 0.24. These results validate the hypothesis that larger receptive fields improve segmentation performance by capturing broader anatomical context. Notably, ${M^5}_{Base}$ and ${M^5}_{Base} + {M^5}_{Focal}$ surpassed the baseline Dice score and achieved a lower NormHD, despite employing substantially fewer parameters per model (32.1M vs. 140M per model instance) and using only two models in the final ensemble compared to the five-model nnU-Net baseline. This demonstrates that improved architectural efficiency and targeted loss function strategies can deliver state-of-the-art performance while reducing computational requirements.

For pathological complete response (pCR) classification, the self-normalizing network trained on radiomic features achieved a balanced accuracy of 57% on the unseen validation set. Output inspection indicates a tendency toward over-confident binary predictions with minimal probabilistic spread, pointing to a calibration issue. Post-hoc temperature scaling is currently being explored as a corrective measure to soften decision boundaries and produce more reliable probability estimates.

Subgroup analysis revealed performance disparities across fairness variables. Balanced accuracy ranged from 75% in the 51–60 age group to 30% in patients over 71, and from 75% in breast density group 'a' to 41.7% in density group 'd'. These gaps highlight the need for future work focused on improving model generalization across diverse demographic and physiological subgroups, potentially through the integration of radiomic features with relevant clinical variables such as age, menopausal status, and breast density.

While our approach focused on using the first post-contrast phase for radiomic feature extraction, prior studies suggest that incorporating temporal kinetics, such as subtracting the pre-contrast from the post-contrast phase can improve the discrimination of tumor response by enhancing contrast uptake patterns [25]. Future work may explore the utility of such subtraction imaging, as well as systematically evaluating other individual post-contrast phases to determine which time point best preserves discriminative features for pCR prediction, and benchmarking other model types alongside the SNN.

Additional efforts will focus on integrating calibration techniques and fairness-aware optimization to improve robustness and reduce subgroup performance disparities in both segmentation and pCR classification.

5 Conclusion

This work presented a large-kernel MedNeXt architecture, combined with the UpKern strategy, for breast tumor segmentation on the MAMA-MIA DCE-MRI dataset. Expanding the kernels from $3 \times 3 \times 3$ to $5 \times 5 \times 5$ improved segmentation performance on the unseen validation set, confirming the benefit of larger receptive fields. For pCR classification, a self-normalizing network trained on radiomic features achieved a balanced accuracy of 57%. The results highlight the need for better calibration and generalization, particularly across demographic subgroups. Future work will explore advanced ensembling and the integration of clinical variables such as age, menopausal status, and breast density to address these challenges.

References

1. Alamoodi, M.: Factors affecting pathological complete response in locally advanced breast cancer cases receiving neoadjuvant therapy: a comprehensive literature review. Eur. J. Breast Health **20**(1), 8 (2023)
2. Askaripour, K., Zak, A.: Breast MRI segmentation by deep learning: key gaps and challenges. IEEE Access **11**, 117935–117946 (2023)
3. Bai, Y., Mei, S., Wang, H., Xiong, C.: Don't just blame over-parametrization for over-confidence: theoretical analysis of calibration in binary classification. In: International Conference on Machine Learning, pp. 566–576. PMLR (2021)
4. Braman, N.M., et al.: Intratumoral and peritumoral radiomics for the pretreatment prediction of pathological complete response to neoadjuvant chemotherapy based on breast DCE-MRI. Breast Cancer Res. **19**, 1–14 (2017)
5. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9901, pp. 424–432. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46723-8_49
6. Debbi, K., et al.: Radiomics model to classify mammary masses using breast DCE-MRI compared to the BI-RADS classification performance. Insights Imaging **14**(1), 64 (2023)

7. Ding, X., Zhang, X., Han, J., Ding, G.: Scaling up your kernels to 31x31: revisiting large kernel design in CNNs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11963–11975 (2022)
8. Dosovitskiy, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
9. Garrucho, L., et al.: A large-scale multicenter breast cancer DCE-MRI benchmark dataset with expert segmentations. Sci. Data **12**(1), 453 (2025)
10. Goorts, B., et al.: MRI-based response patterns during neoadjuvant chemotherapy can predict pathological (complete) response in patients with breast cancer. Breast Cancer Res. **20**(1), 34 (2018)
11. Hortobagyi, G.N., et al.: The global breast cancer burden: variations in epidemiology and survival. Clin. Breast Cancer **6**(5), 391–401 (2005)
12. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat. Methods **18**(2), 203–211 (2021)
13. Klambauer, G., Unterthiner, T., Mayr, A., Hochreiter, S.: Self-normalizing neural networks. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
14. Liu, Z., et al.: Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022 (2021)
15. Lobbes, M.B.I., et al.: The role of magnetic resonance imaging in assessing residual disease and pathologic complete response in breast cancer patients receiving neoadjuvant chemotherapy: a systematic review. Insights Imaging **4**(2), 163–175 (2013). https://doi.org/10.1007/s13244-013-0219-y
16. Milletari, F., Navab, N., Ahmadi, S.A.: V-Net: fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV), pp. 565–571. IEEE (2016)
17. O'Donnell, J., et al.: The accuracy of breast MRI radiomic methodologies in predicting pathological complete response to neoadjuvant chemotherapy: a systematic review and network meta-analysis. Eur. J. Radiol. **157**, 110561 (2022)
18. Oktay, O., et al.: Attention U-Net: learning where to look for the pancreas. arXiv preprint arXiv:1804.03999 (2018)
19. Radhakrishna, S., et al.: Role of magnetic resonance imaging in breast cancer management. South Asian J. Cancer **7**(2), 69 (2018)
20. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
21. Roy, S., et al.: MedNeXt: transformer-driven scaling of convnets for medical image segmentation. In: Greenspan, H., et al. (eds.) MICCAI 2023. LNCS, vol. 14223, pp. 405–415. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-43901-8_39
22. Schott, A.F., et al.: Clinical and radiologic assessments to predict breast cancer pathologic complete response to neoadjuvant chemotherapy. Breast Cancer Res. Treat. **92**(3), 231–238 (2005)
23. Shao, G., Fan, L., Zhang, J., Dai, G., Xie, T.: Association of DW/DCE-MRI features with prognostic factors in breast cancer. Int. J. Biol. Markers **32**(1), 118–125 (2017)
24. Griethuysen, J.J., et al.: Computational radiomics system to decode the radiographic phenotype. Can. Res. **77**(21), e104–e107 (2017)

25. Zhang, X., Teng, X., Zhang, J., Lai, Q., Cai, J.: Enhancing pathological complete response prediction in breast cancer: the role of dynamic characterization of DCE-MRI and its association with tumor heterogeneity. Breast Cancer Res. **26**(1), 77 (2024)
26. Zhu, H., Chen, B., Yang, C.: Understanding why VIT trains badly on small datasets: an intuitive perspective. arXiv preprint arXiv:2302.03751 (2023)



Towards Affordable Tumor Segmentation and Visualization for 3D Breast MRI Using SAM2

Solha Kang[1,2], Eugene Kim[1], Joris Vankerschaver[1,2], and Utku Ozbulak[1,3](✉)

[1] Center for Biosystems and Biotech Data Science, Ghent University Global Campus, Incheon, Republic of Korea
utku.ozbulak@ghent.ac.kr

[2] Department of Mathematics, Computer Science and Statistics, Ghent University, Ghent, Belgium

[3] IDLab, Department of Electronics and Information Systems, Ghent University, Ghent, Belgium

Abstract. Breast MRI provides high-resolution volumetric imaging critical for tumor assessment and treatment planning, yet manual interpretation of 3D scans remains labor-intensive and subjective. While AI-powered tools hold promise for accelerating medical image analysis, adoption of commercial medical AI products remains limited in low- and middle-income countries due to high license costs, proprietary software, and infrastructure demands. In this work, we investigate whether the Segment Anything Model 2 (SAM2) can be adapted for low-cost, minimal-input 3D tumor segmentation in breast MRI. Using a single bounding box annotation on one slice, we propagate segmentation predictions across the 3D volume using three different slice-wise tracking strategies: top-to-bottom, bottom-to-top, and center-outward. We evaluate these strategies across a large cohort of patients and find that center-outward propagation yields the most consistent and accurate segmentations. Despite being a zero-shot model not trained for volumetric medical data, SAM2 achieves strong segmentation performance under minimal supervision. We further analyze how segmentation performance relates to tumor size, location, and shape, identifying key failure modes. Our results suggest that general-purpose foundation models such as SAM2 can support 3D medical image analysis with minimal supervision, offering an accessible and affordable alternative for resource-constrained settings.

Keywords: Equitable AI · Breast Cancer · Tumor Segmentation · SAM2

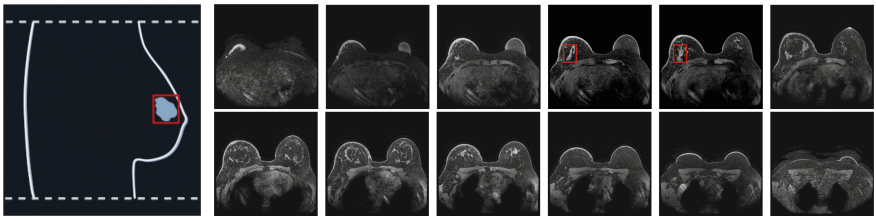
1 Introduction

Breast cancer is the most commonly diagnosed cancer worldwide and is a leading cause of cancer-related mortality, particularly among women [16,23]. Early

S. Kang and E. Kim—Equal contribution.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
T. Zhang et al. (Eds.): Deep-Breath 2025, LNCS 16142, pp. 81–90, 2026.
https://doi.org/10.1007/978-3-032-05559-0_9

and accurate detection is critical for improving patient outcomes, and magnetic resonance imaging (MRI) plays a vital role in this process. In clinical settings, 3D breast MRIs provide high-resolution, volumetric information that can reveal subtle morphological features of tumors [6]. This makes them particularly useful in assessing dense breast tissue, evaluating tumor extent, and planning treatment. However, these volumetric scans often consist of hundreds of axial slices (see Fig. 1), requiring clinicians to mentally reconstruct and interpret the full 3D context, a task that is time-consuming and prone to subjectivity [18,19].



(a) Side profile (b) Example MRI slices

Fig. 1. (a) A side profile diagram of the breast, highlighting the imaging region bounded by two dashed lines. The tumor is illustrated as a blue mass, with a red box indicating its location. (b) Example MRI slices obtained from a cross-sectional region of the breast, shown in sequential order. The slice positions correspond to the imaging region illustrated in (a), and the red box highlights the tumor for slices intersecting the 3D bounding box annotation. (Color figure online)

At the same time, the field of computer vision and artificial intelligence has seen rapid advancements in recent years, with powerful models and pretrained architectures becoming increasingly available through open-source platforms [2,9,25]. Techniques that once required significant technical resources are now widely accessible to researchers and developers. Despite this progress, the integration of AI into real-world healthcare systems, particularly in under-resourced settings, remains limited [1,13]. Proprietary software, licensing fees, and infrastructure requirements often create a financial barrier, rendering AI-powered diagnostic tools inaccessible to many hospitals in low- and middle-income countries [7,20].

On the side of research, recent approaches to 3D breast tumor segmentation have primarily relied on supervised deep learning methods, including 3D convolutional neural networks (CNNs), U-Net variants [11,24], and transformer-based architectures [3,10]. These models typically require large annotated datasets and intensive training efforts, which limit their employability in many scenarios. While these methods have shown strong performance on benchmark datasets, their dependence on dense voxel-wise annotations and task-specific tuning presents a barrier to generalization and scalability, particularly in low-data or low-resource scenarios.

In this work, we examine how the recent advances in video object tracking [4] can be harnessed to develop accessible, low-cost AI solutions for 3D breast MRI interpretation, with a focus on tools that require minimal infrastructure (i.e., no training or fine-tuning). Beyond technical performance, our work emphasizes the importance of equitable access in medical AI. Building entirely on open-source tools and demonstrating their viability for 3D breast MRI segmentation and visualization, we advocate for affordable AI solutions that can be adopted across diverse healthcare systems.

To concretely realize this vision of accessible AI and following the works of [17,26] we leverage recent advances in vision foundation models, specifically employing the open-source SAM2 model [21], to segment and visualize breast tumors in 3D MRI volumes using only minimal human-in-the-loop input (a single bounding box). We show that, although not originally designed for volumetric data, SAM2 can effectively segment breast tumors across slices and generate clinically meaningful visualizations. At the same time, we identify and analyze failure modes in this setting, offering a critical examination of when and why the model breaks down. Our goal is to demonstrate that such tools can deliver practical and accurate results without relying on expensive commercial software or high-end infrastructure, making them viable for deployment in resource-constrained healthcare environments.

2 Methodology

2.1 Dataset

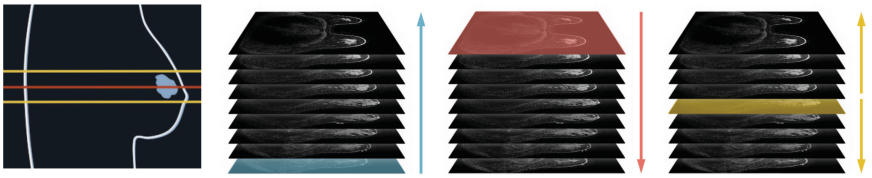
The Duke Breast Cancer Dataset is a large-scale collection of pre-operative 3D breast MRI scans curated from 922 patients treated at Duke University Medical Center [22]. The dataset includes 3D bounding box annotations of tumor locations for each cancer patient and is frequently used as a benchmark for detecting tumors in 2D MRI slices. For our task, we follow the procedure employed in prior work to extract 2D horizontal slice images from the pre-contrast DCE-MRI volumes [5,12,14,15]. This process results in over 140,000 2D images, averaging approximately 130 slices per patient. Among these, tumor-positive slices occur in about 27 images per patient, totaling more than 20,000 images across the dataset.

In 2025, an expanded version of the dataset, referred to as the MAMA-MIA dataset, was released, featuring voxel-level tumor segmentations verified by expert reviewers [8]. This version includes approximately 9,000 segmentation masks corresponding to tumor-positive MRI slices, averaging about 34 masks per patient. In our study, we use the labeled images from MAMA-MIA dataset for 279 patients.

2.2 Model

SAM2 is an open-source foundation model built on vision transformers for the purpose of zero-shot image and video segmentation [21]. It accepts prompts

such as points, bounding boxes, or masks, and produces segmentation masks without requiring task-specific training. In this work, we leverage its tracking functionality to propagate a single-slice bounding box prompt across volumetric breast MRI data for 3D tumor segmentation.



(a) Side profile (b) Image propagation

Fig. 2. (a) A side profile diagram of the breast, highlighting tumor-positive slices bounded by two yellow lines. The center slice of the tumor-positive region is depicted with a red line. (b) Visualization of propagation strategies employed in this work, each illustrated by the arrows from their respective starting slice. (Color figure online)

2.3 3D Tumor Segmentation

To perform 3D tumor segmentation, we treat the volumetric breast MRI scan as a sequence of 2D axial slices. SAM2 is applied in a slice-wise fashion, where a segmentation mask is generated for each slice using the model's tracking functionality. The process begins with a single bounding box prompt provided on one slice, which serves as the initialization for the segmentation. From this initial input, SAM2 tracks the object across adjacent slices by using the predicted mask from the previous slice as contextual guidance for the next.

We evaluate three propagation strategies for traversing the 3D volume below and illustrate them in Fig. 2.

Bottom-to-Top. Starting from the bottom-most slice containing the tumor, we provide a bounding box prompt and sequentially propagate the segmentation upward. Each prediction uses the previous mask as a prompt.

Top-to-Bottom. This approach mirrors the bottom-to-top strategy, but starts from the top-most tumor slice and moves downward through the volume.

Center-Outward. Segmentation begins at the central slice of the tumor, which typically exhibits the largest and clearest tumor region. From this central point, propagation proceeds both upward and downward. This strategy aims to take advantage of the most reliable initial mask and reduce tracking errors over long ranges.

Volumetric Dice Similarity Coefficient. To evaluate the performance of our segmentation model, we use the Volumetric Dice Similarity Coefficient, a standard metric for measuring spatial overlap between the predicted segmentation and the ground truth in 3D medical imaging.

Given a predicted binary volume $\mathtt{P} \in \{0,1\}^{D\times W\times H}$ and a corresponding ground truth volume $\mathtt{G} \in \{0,1\}^{D\times W\times H}$, the volumetric Dice score is defined as:

$$\mathrm{Dice}(\mathtt{P},\mathtt{G}) = \frac{2\cdot|\mathtt{P}\cap\mathtt{G}|}{|\mathtt{P}|+|\mathtt{G}|}, \tag{1}$$

where $|\mathtt{P}\cap\mathtt{G}|$ denotes the number of voxels where the prediction and ground truth both label the same voxel as foreground (i.e., true positives), and $|\mathtt{P}|$ and $|\mathtt{G}|$ are the total number of foreground voxels in the prediction and ground truth volumes, respectively.

This formulation aggregates all voxel-level predictions across the entire 3D volume before computing the score, providing a global measure of overlap. The Dice score ranges from 0 (no overlap) to 1 (perfect agreement) and is particularly robust in settings with imbalanced class distributions, such as tumor segmentation where the foreground occupies a small fraction of the scan.

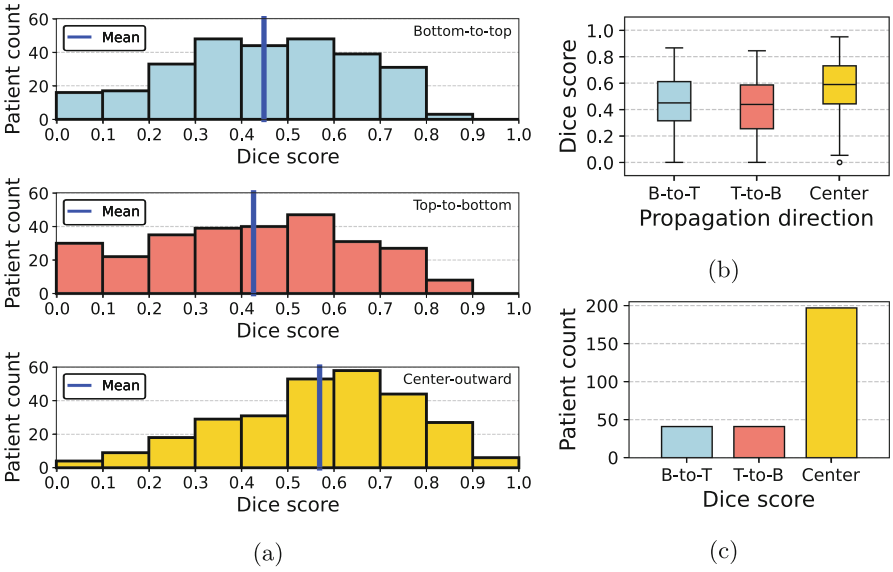


Fig. 3. Evaluation of 3D propagation strategies. (a) Dice score histograms per propagation strategy with mean lines. (b) Boxplots of Dice scores grouped by propagation strategy. (c) Patient counts with highest Dice score per strategy.

3 Experimental Results

Quantitative Results. We evaluate segmentation performance using the setup described in Sect. 2, where a single bounding box is used as input for tracking-based propagation across the 3D volume. As described, we employ three propagation strategies: bottom-to-top, top-to-bottom, and center-outward. The resulting

Dice score distributions across patients are shown in Fig. 3a, with the mean score for each method indicated in blue lines. To compare the overall performance of the three strategies, we further aggregate these results and plot the Dice scores for all patients using boxplots in Fig. 3b. Finally, we present a histogram showing the number of patients for which each propagation strategy achieved the highest Dice score in Fig. 3c, providing insight into the consistency of each method across the dataset.

Among the three propagation strategies, the center-outward method outperforms the bottom-to-top and top-to-bottom approaches, showing higher median Dice scores in Fig. 3b and a more favorable right-skewed distribution in Fig. 3a. This trend is further confirmed in Fig. 3c, where the center-outward method yields the best performance for the majority of patients. These findings suggest that initializing segmentation from the central slice provides a more stable and effective propagation path. This is likely because the tumor tends to be largest and most clearly defined in the center, making it easier for the model to track its boundaries outward compared to starting from peripheral slices. Overall, the segmentation results are encouraging, especially considering that the SAM2 model was not explicitly trained for volumetric breast MRI data and that the only input provided is a single-slice bounding box mask. The ability to achieve robust 3D segmentations under such minimal supervision highlights the potential of leveraging foundation models such as SAM2 in data-scarce or low-resource clinical settings.

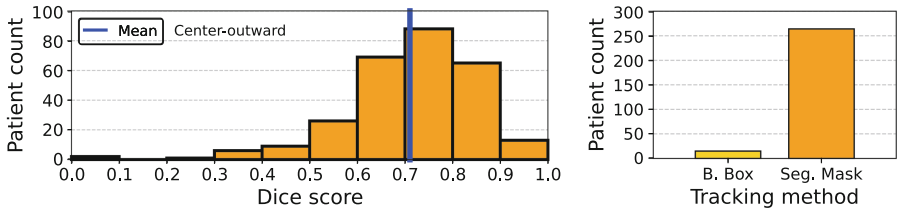
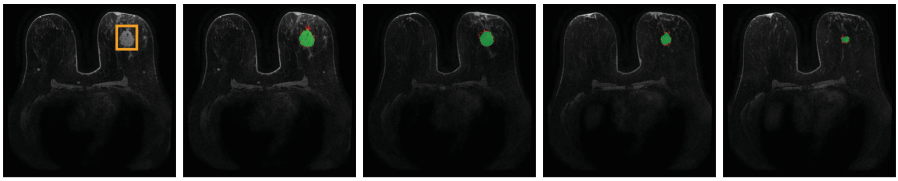


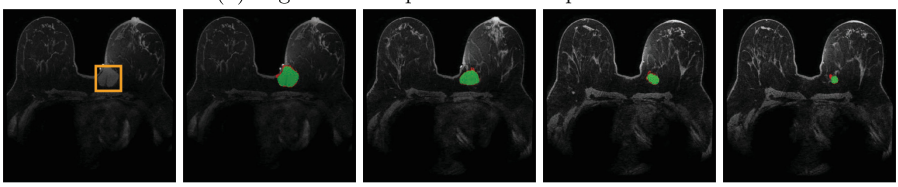
Fig. 4. Evaluation of tumor tracking methods for SAM2. The histogram (left) shows the distribution of Dice scores across patients. The bar plot (right) compares the number of patients with higher Dice scores when using bounding boxes versus segmentation masks as prompts for SAM2 tracking.

Comparison to Mask-Based Tracking. To compare the performance of bounding box tracking with segmentation mask tracking, we rerun the center-outward propagation experiment by providing a segmentation mask on a single central slice as input, as opposed to bounding box. From this initial mask, the model tracks and propagates the segmentation to the remaining slices, using each predicted mask to guide the next step. These results are shown in Fig. 4.

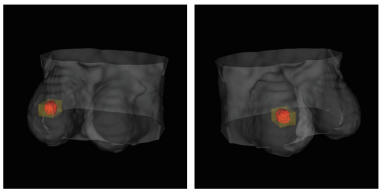
For nearly all patients, segmentation accuracy improves with mask-based guidance compared to bounding box tracking, with only a few exceptions. This mask-based tracking provides a more precise and context-aware prior than the



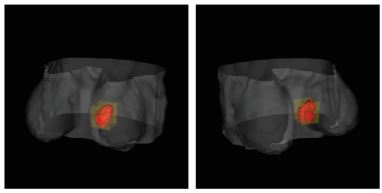
(a) Segmentation predictions for patient 350



(b) Segmentation predictions for patient 400



(c) 3D view for patient 350



(d) 3D view for patient 400

Fig. 5. Sequential MRI slices for (a) Patient 350 and (b) Patient 400. The leftmost slice shows the input with an orange bounding box indicating the tumor region. The next four slices show SAM2 tracking results (center-outward), where green denotes overlap between prediction and ground truth, and red indicates disagreement. (c) and (d) show 3D visualizations of the predicted tumor (red) within the input bounding box (orange). (Color figure online)

coarse bounding box approach. As a result, we observe a substantial improvement in segmentation quality, with the mean Dice score across all patients increasing from 0.57 (bounding box tracking) to 0.71 (segmentation mask tracking). This improvement highlights the benefit of using more informative guidance during propagation, particularly in complex anatomical regions where bounding boxes may fail to represent tumor boundaries accurately, further validating the potential of SAM2-based workflows under minimal supervision.

Qualitative Results. To illustrate the segmentation performance qualitatively, we present example predictions in Fig. 5, including selected 2D slices and 3D reconstructions that highlight the predicted tumor regions. These visualizations demonstrate that the model is able to produce coherent and anatomically plausible segmentations across slices. Additional 3D visualizations, which further support the consistency and quality of the predictions, are included in the supplementary material.

Factors Affecting Dice Score. To explore factors associated with successful and unsuccessful segmentations, we analyze the correlation between the volu-

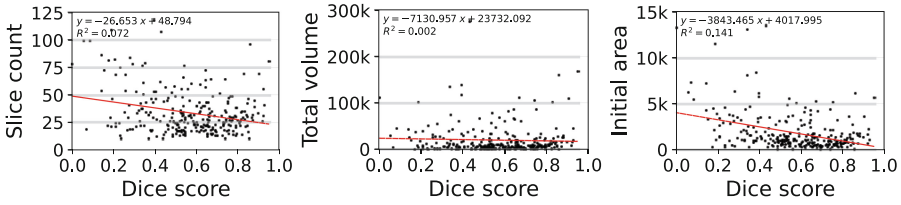


Fig. 6. Scatter plots showing relationships between Dice score and tumor properties. Each plot shows Dice score on the x-axis and a tumor-related variable on the y-axis: (left) number of slices containing tumor, (middle) total tumor volume in voxels, (right) initial tumor area in the tracking slice. A linear regression line is fitted in red, with corresponding equation and R^2 value displayed in each subplot. (Color figure online)

metric Dice score and three variables: (1) the number of slices containing tumor, (2) the total tumor volume, and (3) the initial tumor area in the tracking frame. The results are shown in Fig. 6. As the fitted trend lines and low R^2 values indicate, none of these variables show a meaningful correlation with segmentation performance.

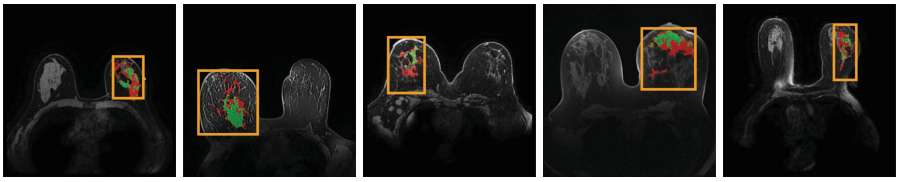


Fig. 7. Example cases where SAM2 yields low Dice scores due to the presence of multiple small lesions, rather than a single large, well-defined tumor.

4 Conclusion and Future Perspectives

In this study, we demonstrate the effective adaptation of SAM2 for invasive MRI breast tumor segmentation with minimal human input. Evaluating three propagation strategies, we found that center-outward initialization results in more accurate and stable segmentation output compared to the other two approaches. These results, achieved with only a single bounding box annotation per patient, highlight the potential of vision foundation models to perform robust volumetric medical imaging segmentation, which they were not originally designed or trained for.

As discussed in Sect. 3, we investigated a number of properties of MRIs to discover cases where SAM2 fails to segment tumors accurately. Our analysis did not reveal any strong correlation between segmentation performance and factors such as tumor volume, number of affected slices, or initial tumor area. However,

upon conducting a qualitative analysis, we discovered that segmentation failures commonly occurred in cases with multiple small lesions scattered across slices, rather than a single large, well-defined tumor (see Fig. 7). These cases challenge the model's ability to maintain spatial coherence during propagation. In future work, we believe that incorporating lightweight pre-selection strategies to better handle fragmented tumor presentations, or integrating uncertainty-aware mechanisms, could further improve segmentation robustness under minimal supervision.

Our work opens up several promising directions for future research. One natural extension is to evaluate and compare the performance of other segmentation foundation models, such as MedSAM2 [17], which are specifically adapted for medical images and volumetric data. Furthermore, combining SAM2 with lightweight pre-processing modules for lesion detection or integrating temporal consistency mechanisms may enhance segmentation robustness in complex cases. Exploring domain adaptation strategies or fine-tuning with a small amount of annotated data could also improve performance while maintaining accessibility in low-resource settings.

References

1. Alowais, S.A., et al.: Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC Med. Educ. **23**(1), 689 (2023)
2. Awais, M., et al.: Foundational models defining a new era in vision: a survey and outlook. arXiv preprint arXiv:2307.13721 (2023)
3. Chen, J., Lu, Y., Yu, Q., et al.: TransUNet: transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)
4. Cheng, H.K., Schwing, A.G.: XMem: long-term video object segmentation with an Atkinson-Shiffrin memory model. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) ECCV 2022. LNCS, vol. 13688, pp. 640–658. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-19815-1_37
5. Chung, M., Won, J.B., Kim, G., Kim, Y., Ozbulak, U.: Evaluating visual explanations of attention maps for transformer-based medical imaging. In: Celebi, M.E., Reyes, M., Chen, Z., Li, X. (eds.) MICCAI 2024. LNCS, vol. 15274, pp. 110–120. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-77610-6_11
6. Clift, A.K., et al.: The current status of risk-stratified breast screening. Br. J. Cancer **126**(4), 533–550 (2022)
7. Dangi, R.R., Sharma, A., Vageriya, V.: Transforming healthcare in low-resource settings with artificial intelligence: recent developments and outcomes. Public Health Nurs. **42**(2), 1017–1030 (2025)
8. Garrucho, L., et al.: A large-scale multicenter breast cancer DCE-MRI benchmark dataset with expert segmentations. Sci. Data **12**(1), 453 (2025)
9. Han, X., et al.: Pre-trained models: past, present and future. AI Open **2**, 225–250 (2021)
10. Hatamizadeh, A., Yin, H., Molchanov, P., et al.: UNETR: transformers for 3D medical image segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 574–584 (2022)

11. Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J.P., Maier-Hein, K.H.: nnU-Net: self-adapting framework for U-Net-based medical image segmentation. Nat. Methods **18**(2), 203–211 (2021)
12. Kang, S., De Neve, W., Rameau, F., Ozbulak, U.: Exploring patient data requirements in training effective AI models for MRI-based breast cancer classification. In: Mann, R.M., et al. (eds.) Artificial Intelligence and Imaging for Diagnostic and Treatment Challenges in Breast Care. Deep-Breath 2024. LNCS, vol. 15451, pp. 75–84. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-77789-9_8
13. Kelly, C.J., Karthikesalingam, A., Suleyman, M., Corrado, G., King, D.: Key challenges for delivering clinical impact with artificial intelligence. BMC Med. **17**, 1–9 (2019)
14. Konz, N., Gu, H., Dong, H., Mazurowski, M.A.: The intrinsic manifolds of radiological images and their role in deep learning. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022. LNCS, vol. 13438, pp. 684–694. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16452-1_65
15. Konz, N., Mazurowski, M.A.: The effect of intrinsic dataset properties on generalization: unraveling learning differences between natural and medical images. arXiv preprint arXiv:2401.08865 (2024)
16. Łukasiewicz, S., Czeczelewski, M., Forma, A., Baj, J., Sitarz, R., Stanisławek, A.: Breast cancer - epidemiology, risk factors, classification, prognostic markers, and current treatment strategies - an updated review. Cancers **13**(17), 4287 (2021)
17. Ma, J., et al.: MedSAM2: segment anything in 3D medical images and videos. arXiv preprint arXiv:2504.03600 (2025)
18. Mango, V.L., et al.: Abbreviated protocol for breast MRI: are multiple sequences needed for cancer detection? Eur. J. Radiol. **84**(1), 65–70 (2015)
19. Mootz, A.R., Madhuranthakam, A.J., Doğan, B.: Changing paradigms in breast cancer screening: abbreviated breast MRI. Eur. J. Breast Health **15**(1), 1 (2019)
20. Panch, T., Mattie, H., Celi, L.A.: The "inconvenient truth" about AI in healthcare. NPJ Digit. Med. **2**(1), 1–3 (2019)
21. Ravi, N., et al.: SAM2: segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)
22. Saha, A., et al.: A machine learning approach to radiogenomics of breast cancer: a study of 922 subjects and 529 DCE-MRI features. Br. J. Cancer **119**(4), 508–516 (2018)
23. Siegel, R.L., Giaquinto, A.N., Jemal, A.: Cancer statistics, 2024. CA Cancer J. Clin. **74**(1), 12–49 (2024)
24. Zhang, Z., et al.: UniMRISegNet: universal 3D network for various organs and cancers segmentation on multi-sequence MRI. IEEE J. Biomed. Health Inf. (2024)
25. Zhou, C., et al.: A comprehensive survey on pretrained foundation models: a history from BERT to ChatGPT. Int. J. Mach. Learn. Cybern., 1–65 (2024)
26. Zhu, J., Hamdi, A., Qi, Y., Jin, Y., Wu, J.: Medical SAM2: segment medical images as video via segment anything model 2. arXiv preprint arXiv:2408.00874 (2024)



Breast Cancer Detection from Multi-view Screening Mammograms with Visual Prompt Tuning

Han Chen[1,2(✉)] and Anne L. Martel[1,2]

[1] Physical Sciences, Sunnybrook Research Institute, Toronto, ON, Canada
han.chen@sri.utoronto.ca
[2] Medical Biophysics, University of Toronto, Toronto, ON, Canada

Abstract. Accurate detection of breast cancer from high-resolution mammograms is crucial for early diagnosis and effective treatment planning. Previous studies have shown the potential of using single-view mammograms for breast cancer detection. However, incorporating multi-view data can provide more comprehensive insights. Multi-view classification, especially in medical imaging, presents unique challenges, particularly when dealing with large-scale, high-resolution data. In this work, we propose a novel Multi-view Visual Prompt Tuning Network (MVPT-NET) for analyzing multiple screening mammograms. We first pretrain a robust single-view classification model on high-resolution mammograms and then innovatively adapt multi-view feature learning into a task-specific prompt tuning process. This technique selectively tunes a minimal set of trainable parameters (7%) while retaining the robustness of the pre-trained single-view model, enabling efficient integration of multi-view data without the need for aggressive downsampling. Our approach offers an efficient alternative to traditional feature fusion methods, providing a more robust, scalable, and efficient solution for high-resolution mammogram analysis. Experimental results on a large multi-institution dataset demonstrate that our method outperforms conventional approaches while maintaining detection efficiency, achieving an AUROC of 0.852 for distinguishing among Non-cancer, DCIS, and Invasive classes. This work highlights the potential of MVPT-NET for medical imaging tasks and provides a scalable solution for integrating multi-view data in breast cancer detection.

Keywords: Breast cancer detection · Screening mammogram · Vision transformer · Multi-view · Visual prompt tuning

1 Introduction

Breast cancer is a leading cause of cancer-related mortality among women worldwide, representing a significant public health challenge [6]. Early detection through screening mammography is essential for improving patient outcomes and

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
T. Zhang et al. (Eds.): Deep-Breath 2025, LNCS 16142, pp. 91–100, 2026.
https://doi.org/10.1007/978-3-032-05559-0_10

reducing mortality rates. Mammography remains the most reliable tool for early breast cancer diagnosis, typically capturing two high-resolution images of each breast from different perspectives: mediolateral-oblique (MLO) and craniocaudal (CC). Unlike conventional deep learning approaches that primarily rely on single-view analysis, radiologists interpret all views together, drawing on cross-view correlations to improve diagnostic accuracy. This practice illustrates the value of multi-view analysis in identifying abnormalities and supporting clinical decision-making. As a result, multi-view or multi-image-based Computer-Aided Diagnosis (CAD) systems have gained attention for their potential to surpass the performance of single-view models [1,5]. Recent advancements in deep learning have further enhanced breast cancer classification and detection, with increasing interest in multi-view architectures inspired by radiologists' interpretive strategies [12]. These developments suggest that multi-view learning could play a vital role in advancing breast cancer detection and diagnosis.

Recent advances in artificial intelligence, particularly deep neural networks, have demonstrated remarkable capabilities in extracting informative imaging features for CAD [2,4,7]. Among various deep learning architectures [3,14,16,21], Transformers [21] have gained significant attention in medical image analysis. Transformers partition an image into a sequence of small patches and employ multi-head self-attention mechanisms to model dense long-range dependencies. Their self-attention design effectively captures both local and global features, providing an advantage over Convolutional Neural Networks (CNNs) [14]. Despite the success, Transformers suffer from quadratic time complexity relative to sequence length, making them computationally expensive for high-resolution images. This issue is particularly critical in multi-view mammogram analysis, where leveraging correlations across different views can significantly enhance diagnostic accuracy. Although several methods [15,19,20] have been proposed for multi-view mammogram analysis, their performance is often constrained by aggressive downsampling. This downsampling results in a loss of fine-grained details, which are crucial for accurate breast cancer detection. Consequently, current models struggle to fully exploit the rich information contained in high-resolution, multi-view mammograms.

Recently, Visual Prompt Tuning (VPT) [13] has emerged as an efficient method for adapting large pre-trained vision Transformers to new tasks. Originating from prompting [18] in natural language processing, VPT extends the concept of prepending task-specific instructions to input data, enabling models to generalize to new tasks with minimal adjustments. Unlike traditional fine-tuning, which updates all model parameters, VPT introduces a small set of learnable prompts in the input space while keeping the backbone frozen, significantly reducing the number of trainable parameters. Motivated by its efficiency and strong performance in vision tasks, recent works have increasingly applied VPT to medical image processing [10,23].

In this study, we propose a novel Multi-view Visual Prompt Tuning Network (MVPT-NET) for efficient multi-view mammogram classification in breast cancer detection. We treat this task as a two-stage learning process. Firstly,

to leverage the power of pre-trained models, we pretrain a Swin Transformer on a single-view classification task. Secondly, during downstream prompt tuning, a small set of task-specific learnable parameters from multi-view inputs is introduced into the input space, while the entire pre-trained Swin Transformer backbone remains frozen. To further enhance the cross-view correlation learning, tokens from different views are merged and passed through the transformer blocks to obtain the final comprehensive prediction. MVPT-NET sets itself apart from conventional approaches by efficiently utilizing pre-trained single-view models, and tuning with multi-view data, retaining high-resolution information, and avoiding aggressive downsampling. We use data from 4,179 subjects across multiple NHS Breast Cancer screening sites in the UK [9] to evaluate the performance of MVPT-NET for breast cancer classification. Experimental results show that MVPT-NET achieves an Area under the Receiver Operating Characteristic Curve (AUROC) of 0.852 for distinguishing between Non-cancer, DCIS, and Invasive cancers. Notably, only 7% of all model parameters are tuned during training. Our method demonstrates strong performance with a relatively small percentage of parameter adjustments, highlighting its potential for multi-view mammogram analysis in breast cancer detection.

2 Methods

In this section, we describe our baseline model and MVPT-NET (illustrated in Fig. 1). Our proposed method involves two phases of training: single-view baseline model pretraining and downstream multi-view visual prompt tuning with mammogram pairs, along with cross-view knowledge distillation.

2.1 Single-View Pretraining

Our single-view baseline model follows the Swin Transformer Base architecture [16], which is chosen as the backbone for MVPT-NET due to its hierarchical feature representation and efficient modeling of long-range dependencies using shifted window attention. Given an input mammogram $I \in \mathbb{R}^{H \times W \times C}$, where H and W are the height and width of the input image, and C is the number of channels, the image is first divided into fixed-sized patches of size $h \times w$. These patches are then linearly embedded into a latent space, producing patch tokens $\mathrm{E}_0 \in \mathbb{R}^{m \times d}$, where $m = \frac{H}{h} \times \frac{W}{w}$ denotes the number of patches, and d is the embedding dimension:

$$\mathrm{E}_0 = \mathrm{Embed}(I), \quad \mathrm{E}_0 \in \mathbb{R}^{m \times d} \tag{1}$$

The embedded patches are then passed through a series of Swin Transformer layers. Each layer L_i consists of shifted window-based multi-head self-attention (W-MSA) and feed-forward networks (FFN) with residual connections and LayerNorm. The computation in each layer L_i is formulated as:

$$\mathrm{E}_i = L_i(\mathrm{E}_{i-1}), \quad i = 1, 2, \ldots, N \tag{2}$$

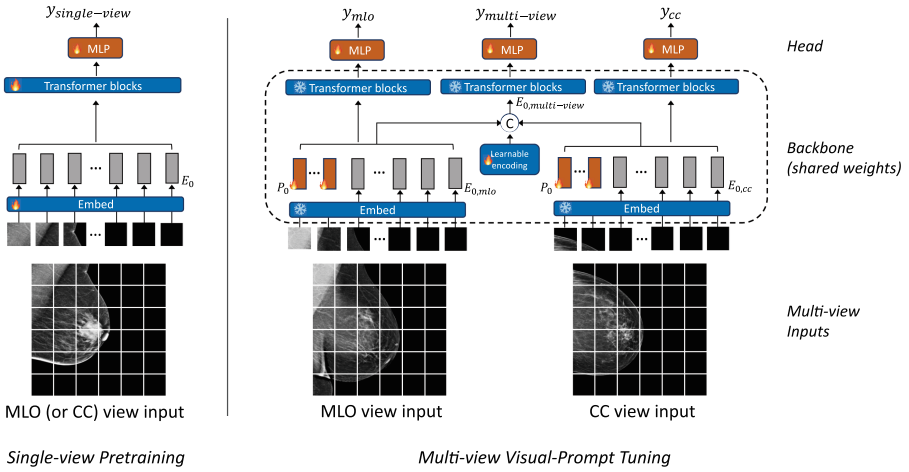


Fig. 1. Network architecture of MVPT-NET

where N is the total number of Swin Transformer layers. Finally, a neural classification head maps the output of the final layer E_N into a predicted class probability distribution y:

$$y = \mathrm{Head}(\mathrm{E}_N) \tag{3}$$

The y is then used to calculate the cross-entropy loss with the single-view mammogram ground truth. The pre-trained Swin Transformer model will serve as the backbone for MVPT-NET, which will be further prompt-tuned in the downstream task of multi-view mammogram classification.

2.2 Multi-view Visual Prompt Tuning

Preliminaries. Visual Prompt Tuning (VPT) [13] introduces a set of p continuous embeddings, or prompts, of dimension d into every Transformer layer's input space after the embedding layer. During tuning, only the task-specific prompts are updated, while the Transformer backbone remains frozen. For the $(i+1)$-th layer L_{i+1}, the collection of learnable prompts is denoted as $\mathrm{P}_i = \{\mathrm{p}_k^i \in \mathbb{R}^d \mid k \in \mathbb{N}, 1 \leq k \leq m\}$. The prompted Swin Transformer model can then be formulated as Eqs. (4), where the symbols $\underline{\ }$ and $\sim$ indicate learnable and frozen parameters, respectively:

$$\begin{aligned} \mathrm{E}_{i,\mathrm{view}} &= \underset{\sim}{L_i}([\underline{\mathrm{P}_{i-1}}, \mathrm{E}_{i-1,\mathrm{view}}]), \quad i = 1, 2, \ldots, N \\ y_{\mathrm{view}} &= \underline{\mathrm{Head}}(\mathrm{E}_N), \quad \mathrm{view} \in \{mlo, cc\} \end{aligned} \tag{4}$$

Multi-view Visual Prompt Tuning. Our proposed MVPT-NET adapts VPT into multi-view mammogram classification. Given paired mammograms as input, the patch embedding layer processes the MLO and CC views respectively, generating tokens denoted as $\mathrm{E}_{0,mlo}$ and $\mathrm{E}_{0,cc}$. Previous studies have explored various fusion strategies for integrating multi-view mammograms in breast cancer detection, such as middle fusion and late fusion. Inspired by the work [1], we adopt an early fusion strategy to effectively combine tokens from multi-view. Specifically, the joint token representation is formulated as:

$$\begin{aligned} \mathrm{E}_{0,joint} &= [\mathrm{P}_0, \mathrm{E}_{0,mlo}; \mathrm{P}_0, \mathrm{E}_{0,cc}] \\ \mathrm{E}_{0,\text{multi-view}} &= \mathrm{E}_{0,joint} + \mathrm{E}_{\text{context}} \end{aligned} \tag{5}$$

where $\mathrm{E}_{\text{context}}$ is a learnable encoding shared across all tokens from each view. This enriched multi-view representation is then fed into the Swin Transformer layers as Eq. (6), enabling the model to capture comprehensive cross-view correlations and get the overall prediction $y_{\text{multi-view}}$ for accurate mammogram classification.

$$\begin{aligned} \mathrm{E}_{i,\text{multi-view}} &= L_i([\mathrm{E}_{i-1,\text{multi-view}}]), \quad i = 1, 2, \ldots, N \\ y_{\text{multi-view}} &= \mathrm{Head}(\mathrm{E}_{N,\text{multi-view}}) \end{aligned} \tag{6}$$

Learning Object and Knowledge Distillation. The overall training loss is the combination of four terms:

$$L_{\text{overall}} = L_{\text{multi-view}} + L_{\text{mlo}} + L_{\text{cc}} + \lambda L_{\text{md}} \tag{7}$$

where the first three terms represent the cross-entropy loss between $y_{\text{multi-view}}$, y_{mlo}, y_{cc} and the ground truth label. λ is a trade-off parameter to balance the contribution of the distillation term L_{md} and the classification terms.

Inspired by the work [1], we incorporate a mutual knowledge distillation loss term L_{md} to encourage collaboration between the y_{mlo}, y_{cc}, and $y_{\text{multi-view}}$. This term is based on the distillation scheme of work [11], which minimizes the Kullback-Leibler divergence between temperature-softened distributions produced by the teacher and student models. The loss is calculated as:

$$L_{\text{kd}}(t, s; \tau) = D_{\text{KL}}(\tilde{\sigma}(t, \tau), \tilde{\sigma}(s, \tau)) \tag{8}$$

The mutual distillation loss L_{md} is formulated as a combination of asymmetric distillation terms between the score-fused class distribution and the multi-view predictions:

$$L_{\text{md}}(\{y_{\text{mlo}}, y_{\text{cc}}\}, y_{\text{multi-view}}; \tau) = \frac{1}{\tau^2}\left[L_{\text{kd}}(\hat{\bar{z}}, y_{\text{multi-view}}; \tau) + L_{\text{kd}}(\hat{y}_{\text{multi-view}}, \bar{z}; \tau)\right] \tag{9}$$

where $\bar{z} = \frac{1}{2}(y_{\text{mlo}} + y_{\text{cc}})$, and the $\hat{\bar{z}}$ represents the gradient-detached copy of $\bar{z}$. This loss term helps align the logits from the multi-view and single-view predictions, improving generalization and performance across different views.

3 Experiments and Results

3.1 Datasets

This study uses a subset of the OPTIMAM dataset [8,9], which includes mammogram images and clinical data collected from four NHS Breast Cancer Screening sites across the UK. All mammograms in this subset were acquired using either Hologic or Siemens systems, with a ratio of approximately 2:1. Subjects are categorized into two label groups based on their clinical records:(1) Non-cancer vs. Cancer and (2) Non-cancer vs. DCIS vs. Invasive. In the first group, a subject is classified as non-cancer if the biopsy confirms a benign result or the screening is normal, and as cancer if the biopsy confirms a malignant result. In the second group, non-cancer subjects are defined in the same way as in the first group. DCIS subjects have biopsy-confirmed malignant results with a recorded DCIS grade but no invasive components. Invasive subjects have biopsy-confirmed malignant results that include invasive components.

The dataset contains 8,358 mammograms from 4,179 subjects, with each subject having two views (MLO and CC). In the first label group, there are 1,500 non-cancer subjects (750 normal and 750 benign) and 2,679 cancer subjects. In the second group, there are 1,500 non-cancer subjects (750 normal and 750 benign), 1,204 DCIS cases, and 1,475 invasive cancer cases. The data is split into a training set (80%, 3,343 subjects) and a test set (20%, 836 subjects), with the training set further divided into five folds for cross-validation. Class proportions are maintained across all splits.

In this study, biopsy-confirmed benign cases and screening-normal cases were grouped under the label "Non-cancer" to simplify the classification task and align with a clinically actionable distinction between malignant and non-malignant findings. A more fine-grained classification that separates benign and normal cases could be explored in future work to better support clinical decision-making.

3.2 Implementation Details

We set the input mammogram size to 1024×1024 and use a Swin Transformer Base [16] architecture with an increased window size (from 7 to 16). For single-view pretraining, we initialized the model with self-supervised pretrained weights from ImageNet [17], adapting them via bicubic interpolation for the larger window size. The initial learning rate was set to 5e−5 with a batch size of 6. We used the AdamW optimizer with a cosine learning rate scheduler, applying a 10-epoch warm-up starting from 1e−6. The model was trained for 50 epochs.

For multi-view visual prompt tuning, we set the base learning rate to 0.01, weight decay to 0.01, and batch size to 2, using the SGD optimizer with a cosine scheduler and a 10-epoch warm-up. The total number of epochs is 100, and the

prompt length is set to 5. For the hyperparameters in L_{md}, we follow the original setup from [1], setting $\tau = 4$ and $\lambda = 0.1$.

Both pretraining and prompt tuning involved data augmentation, including random vertical flipping and affine transformations. Mammograms were reoriented to ensure the breast consistently appears on the left side. Models were trained on an NVIDIA A100 GPU. The evaluation metrics include accuracy, precision, recall, F1-score, and the area under the receiver operating characteristic (AUROC).

3.3 Results of Breast Cancer Detection

We compared our proposed method with several state-of-the-art multi-view classification models, including Multi-View Swin Transformer (MV-Swin-T) [19], Multi-View Chest Radiograph Classification Network (MVC-NET) [22], Cross-View Transformers (CVT) [20], and Multi-view Hybrid Fusion and Mutual Distillation Network (MV-HFMD) [1]. We used the authors' implementations and followed the same training and evaluation process as MVPT-NET. The baseline is trained on both MLO and CC view data but only learns from one view per training sample. The individual predictions for MLO and CC views produced by MVPT-NET are reported as MVPT-NET$_{\text{single-view}}$.

Table 1 presents the Non-cancer vs. Cancer classification results of MVPT-NET compared to several SOTA models. Our method achieves the highest AUROC of 0.842, outperforming MV-Swin-T [19] which uses the same backbone but suffers from a smaller input resolution and learning from multi-view input from scratch. MV-HFMD [1] also uses knowledge distillation but faces learning confusion due to its mutual loss leading to the update of the entire model, resulting in a lower AUROC of 0.786. Additionally, the single-view outputs from MVPT-NET also performs better than the single-view pretrained baseline, demonstrating the advantage of our multi-view visual prompt tuning approach in capturing the full range of information from both the MLO and CC views.

Table 1. Non-cancer vs. Cancer test results. All metrics are reported as mean±standard deviation from 5-fold cross-validation

Method	Accuracy	Precision	Recall	F1-Score	AUROC
MV-Swin-T [19]	0.683 ± 0.020	0.691 ± 0.021	0.660 ± 0.015	0.676 ± 0.025	0.694 ± 0.008
MVC-NET [22]	0.728 ± 0.017	0.733 ± 0.041	0.701 ± 0.030	0.711 ± 0.046	0.736 ± 0.019
CVT [20]	0.708 ± 0.012	0.711 ± 0.018	0.674 ± 0.024	0.704 ± 0.034	0.745 ± 0.016
MV-HFMD [1]	0.725 ± 0.028	0.708 ± 0.028	0.697 ± 0.055	0.710 ± 0.065	0.786 ± 0.011
Baseline (Ours)	0.754 ± 0.014	0.784 ± 0.030	0.743 ± 0.016	0.750 ± 0.021	0.825 ± 0.007
MVPT-NET$_{\text{single-view}}$	0.763 ± 0.007	0.744 ± 0.007	0.746 ± 0.012	0.744 ± 0.007	0.837 ± 0.005
MVPT-NET$_{\text{multi-view}}$	0.774 ± 0.006	0.757 ± 0.008	0.749 ± 0.011	0.750 ± 0.007	0.842 ± 0.005

We also conducted a more challenging fine-grained breast cancer classification task, distinguishing between Non-cancer, Ductal Carcinoma in Situ (DCIS), and Invasive Ductal Carcinoma (Invasive), with the results presented in Table 2. Despite the complexity of the task, our proposed method outperforms all the compared models across all metrics. The success of MVPT-NET can be attributed to pretraining with high-resolution input, avoiding aggressive downsampling, which preserves important details and leads to the learning of a strong backbone. The multi-view visual prompt tuning further learns cross-view features from paired mammogram inputs, updating the prompt without altering the pretrained backbone. This enables the final model to effectively capture both single-view and multi-view information.

Table 2. Non-cancer vs. DCIS vs. Invasive test results. All metrics except accuracy are macro-averaged and reported as mean±standard deviation from 5-fold cross-validation

Method	Accuracy	Precision	Recall	F1-Score	AUROC
MV-Swin-T [19]	0.637 ± 0.023	0.607 ± 0.024	0.638 ± 0.016	0.616 ± 0.028	0.701 ± 0.015
MVC-NET [22]	0.645 ± 0.019	0.612 ± 0.038	0.583 ± 0.031	0.613 ± 0.041	0.730 ± 0.031
CVT [20]	0.651 ± 0.014	0.604 ± 0.024	0.584 ± 0.019	0.600 ± 0.022	0.754 ± 0.013
MV-HFMD [1]	0.636 ± 0.015	0.650 ± 0.047	0.629 ± 0.018	0.651 ± 0.042	0.804 ± 0.010
Baseline (Ours)	0.668 ± 0.010	0.675 ± 0.023	0.664 ± 0.012	0.664 ± 0.011	0.831 ± 0.006
$\text{MVPT-NET}_{\text{single-view}}$	0.673 ± 0.005	0.684 ± 0.007	0.666 ± 0.006	0.666 ± 0.005	0.839 ± 0.005
$\text{MVPT-NET}_{\text{multi-view}}$	0.688 ± 0.007	0.696 ± 0.005	0.681 ± 0.009	0.683 ± 0.009	0.852 ± 0.004

3.4 Ablation Study

We conducted an ablation study on the Non-cancer vs. DCIS vs. Invasive task to assess the impact of L_{md}. Introducing this loss improved AUROC, showing the effectiveness of knowledge distillation in capturing multi-view features (Table. 3).

Table 3. Ablation study results. All metrics except accuracy are macro-averaged and reported as mean±standard deviation from 5-fold cross-validation

Method	Accuracy	Precision	Recall	F1-Score	AUROC
MVPT-NET w/o L_{md}	0.662 ± 0.009	0.688 ± 0.012	0.655 ± 0.014	0.653 ± 0.014	0.849 ± 0.003
MVPT-NET	0.688 ± 0.007	0.696 ± 0.005	0.681 ± 0.009	0.683 ± 0.009	0.852 ± 0.002

4 Conclusions

In conclusion, we proposed MVPT-NET, a novel approach that adapts Visual Prompt Tuning for multi-view mammogram analysis. Our method outperformed several state-of-the-art multi-view models in breast cancer detection, showcasing its ability to leverage multi-view inputs for more accurate predictions. The ablation study highlights the significance of the introduced loss term in improving model performance by effectively capturing and distilling meaningful features from both MLO and CC views. Our approach, with its high-resolution pretraining and multi-view prompt tuning capabilities, offers a promising direction for enhancing breast cancer detection and classification tasks.

References

1. Black, S., Souvenir, R.: Multi-view classification using hybrid fusion and mutual distillation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 270–280 (2024)
2. Chen, H., Jiang, Y., Ko, H., Loew, M.: A teacher-student framework with Fourier transform augmentation for covid-19 infection segmentation in CT images. Biomed. Signal Process. Control **79**, 104250 (2023)
3. Chen, H., Jiang, Y., Loew, M., Ko, H.: Unsupervised domain adaptation based covid-19 CT infection segmentation network. Appl. Intell. **52**(6), 6340–6353 (2022)
4. Chen, H., Martel, A.L.: Towards improved breast cancer detection on digital mammograms using local self-attention-based transformer. In: 17th International Workshop on Breast Imaging (IWBI 2024), vol. 13174, pp. 455–461. SPIE (2024)
5. Chen, Y., et al.: Multi-view local co-occurrence and global consistency learning improve mammogram classification generalisation. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022. LNCS, vol. 13433, pp. 3–13. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16437-8_1
6. Contiero, P., et al.: Causes of death in women with breast cancer: a risks and rates study on a population-based cohort. Front. Oncol. **13**, 1270877 (2023)
7. Gou, F., Liu, J., Xiao, C., Wu, J.: Research on artificial-intelligence-assisted medicine: a survey on medical artificial intelligence. Diagnostics **14**(14), 1472 (2024)
8. Halling-Brown, M.D.: OMI-DB (2020). https://medphys.royalsurrey.nhs.uk/omidb/
9. Halling-Brown, M.D., et al.: OPTIMAM mammography image database: a large-scale resource of mammography images and clinical data. Radiol. Artif. Intell. **3**(1), e200103 (2020)
10. He, A., Wu, Y., Wang, Z., Li, T., Fu, H.: DVPT: dynamic visual prompt tuning of large pre-trained models for medical image analysis. Neural Netw. **185**, 107168 (2025)
11. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
12. Jain, K., Rangarajan, K., Arora, C.: Follow the radiologist: Clinically relevant multi-view cues for breast cancer detection from mammograms. In: Linguraru, M.G., et al. (eds.) MICCAI 2024. LNCS, vol. 15001, pp. 102–112. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-72378-0_10

13. Jia, M., et al.: Visual prompt tuning. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) ECCV 2022. LNCS, vol. 13693, pp. 709–727. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-19827-4_41
14. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. Commun. ACM **60**(6), 84–90 (2017)
15. Li, C., et al.: Multi-view mammographic density classification by dilated and attention-guided residual learning. IEEE/ACM Trans. Comput. Biol. Bioinf. **18**(3), 1003–1013 (2020)
16. Liu, Z., et al.: Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022 (2021)
17. Russakovsky, O., et al.: ImageNet large scale visual recognition challenge. Int. J. Comput. Vision **115**, 211–252 (2015)
18. Sahoo, P., Singh, A.K., Saha, S., Jain, V., Mondal, S., Chadha, A.: A systematic survey of prompt engineering in large language models: techniques and applications. arXiv preprint arXiv:2402.07927 (2024)
19. Sarker, S., Sarker, P., Bebis, G., Tavakkoli, A.: MV-Swin-T: mammogram classification with multi-view Swin transformer. In: 2024 IEEE International Symposium on Biomedical Imaging (ISBI), pp. 1–5. IEEE (2024)
20. Tulder, G., Tong, Y., Marchiori, E.: Multi-view analysis of unregistered medical images using cross-view transformers. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12903, pp. 104–113. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87199-4_10
21. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
22. Zhu, X., Feng, Q.: MVC-Net: multi-view chest radiograph classification network with deep fusion. In: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI), pp. 554–558. IEEE (2021)
23. Zu, W., Xie, S., Zhao, Q., Li, G., Ma, L.: Embedded prompt tuning: towards enhanced calibration of pretrained models for medical images. Med. Image Anal. **97**, 103258 (2024)

Are Vision Foundation Models Ready for Out-of-the-Box Medical Image Registration?

Hanxue Gu[1(✉)], Yaqian Chen[1], Nicholas Konz[1], Qihang Li[2], and Maciej A. Mazurowski[3]

[1] Department of Electrical and Computer Engineering, Duke University, Durham, UK
hanxue.gu@duke.edu
[2] Department of Biostatistics and Bioinformatics, Duke University, Durham, UK
[3] Departments of Biostatistics and Bioinformatics, Radiology, Electrical and Computer Engineering, and Computer Science, Duke University, Durham, UK

Abstract. Vision foundation models (VFM), pre-trained on large image datasets and capable of capturing rich feature representations, have recently shown potential for zero-shot image registration. However, their performance has mostly been tested in the context of less complex structures, such as the brain or abdominal organs, and it remains unclear whether these models can handle more challenging, deformable anatomy. Breast MRI registration is particularly challenging due to anatomical variation between patients, deformation from positioning, and the thin, complex structure of fibroglandular tissue (FGT), where accurate alignment is crucial. In this study, we provide a comprehensive evaluation of VFM-based registration algorithms. We assess five pre-trained encoders, including DINO-v2, SAM, MedSAM, SSLSAM, and MedCLIP, across **four** key breast registration tasks that capture variations in different years and dates, sequences, modalities, and patient disease status (lesion versus no lesion). Our results show that among the five pre-trained encoders, SAM demonstrates a clear advantage. VFM outperforms traditional optimization-based methods, especially in cross-modality tasks, with its advantage more pronounced for large structures (e.g., organs and breast) than for finer anatomical details like FGT. Further work is needed to understand how domain-specific training influences registration and to explore targeted strategies that improve fine structure accuracy. We also publicly release our code at Github.

Keywords: Foundation model · Breast registration · Zero-shot

1 Introduction

Accurate alignment of breast MRI images across longitudinal studies, sequences, or even modalities (*e.g.*,, PET-CT) is essential for tracking tumor changes and

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
T. Zhang et al. (Eds.): Deep-Breath 2025, LNCS 16142, pp. 101–112, 2026.
https://doi.org/10.1007/978-3-032-05559-0_11

assisting surgery planning [1]. However, deformable registration of the breast remains highly challenging due to the natural variability and high deformability of both the breast and dense tissue [6,19]. Most studies still rely on traditional optimization-based registration algorithms [21], such as the symmetric image normalization method (SyN) [2] and Elastix [14]. Several deep learning-based registration methods have recently been proposed, but they typically require large, task-specific training datasets or annotated masks, making them less practical or transferable across different clinical settings [3,5,11].

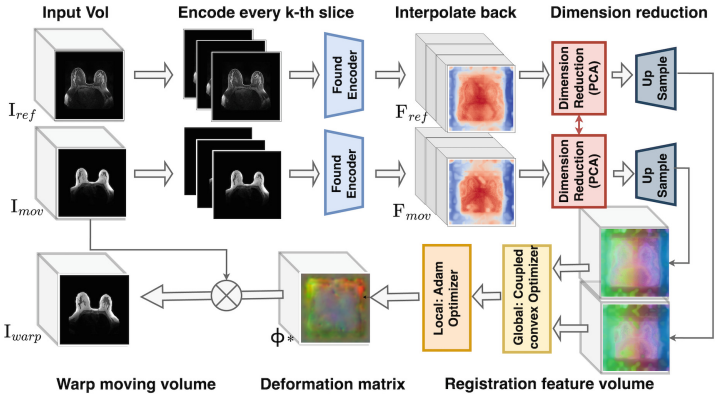


Fig. 1. Pipeline for VFM-based Breast Image Registration. The Found Encoder can be replaced with any pretrained vision foundation encoder

In contrast, vision foundation models (VFMs) have gained attention for their strong generalization ability and zero-shot performance on downstream tasks, such as segmentation and classification [9,18,22]. One recent effort, DINO-Reg [24], shows that DINO-v2's rich semantic embeddings can support effective zero-shot registration across medical imaging domains. Yet, it remains unclear how different pre-training strategies, such as masked autoencoders (MAE), vision-language models (VLMs), or medical domain-specific models like MedSAM, affect registration performance. Moreover, most studies focus on less deformable regions like abdominal organs with larger, more stable structures, while breast registration poses unique challenges with significant shape variation and thin, highly deformable tissue. How VFM-based algorithms handle such complex anatomy remains poorly understood.

To mitigate this gap, we present the first systematic benchmark of various foundation model encoders for zero-shot breast MRI registration. Our study evaluates different types of pretraining strategies, compares foundation-based methods with other zero-shot approaches, and examines the impact of domain-specific versus natural image pretraining. We conduct comprehensive experiments on a curated breast MRI dataset across multiple challenging scenarios, including multi-time same-sequence registration, multi-sequence registration, cross-modality alignment, and lesion tracking.

The main contributions of our work are:

1. We implement a flexible pipeline for foundation-model-based algorithms supporting different pretrained encoders, and make the code publicly available.
2. We conduct a comprehensive comparison of registration strategies according to the pretraining strategies that were applied to develop the foundation model, including contrastive learning, masked autoencoder, and vision-language approaches, across both natural and medical imaging domains.
3. We design four registration tasks that reflect real-world challenges, covering longitudinal, cross-sequence, lesion-tracking, and PET-to-MRI alignment.
4. We assess model performance across targets of varying size, anatomical detail, and deformation, including breast, dense tissue, organ, and lesion.
5. We systematically evaluate the strengths and limitations of VFM-based algorithms compared to other zero-shot approaches.

2 Method

We adopt a training-free registration pipeline that extracts semantic embeddings from input volumes using a pretrained foundation model encoder, followed by deformable registration performed on the embeddings (dimension-reduced) without any additional training or fine-tuning (Sect. 2.1). The pre-trained encoders can be flexibly used as plug-in components. These encoders vary in pre-training strategies (e.g., contrastive learning or masked autoencoding) and in whether they were pre-trained on medical images or natural images (Sect. 2.2).

2.1 VFM-Based Registration

Our training-free, VFM-based registration pipeline, inspired by the framework of [24], includes three main stages: (1) feature extraction, (2) dimension reduction, and (3) feature-based registration, as shown in Fig. 1.

Feature Extraction. Let I_{ref} and I_{mov} denote reference and moving image volumes for registration. Considering that most available pre-trained encoders are for 2D images, we perform feature extraction slice-by-slice along the axial plane. Each volume is treated as a sequence of 2D axial slices, denoted $\{I^z_{\text{ref}}\}_{z=1}^{Z}$ and $\{I^z_{\text{mov}}\}_{z=1}^{Z}$, where z indexes the slice number. To reduce computational cost, we chose to encode every k-th slice, where k is a hyperparameter that can be selected based on the desired speed of registration. The selected slices are passed through a frozen vision encoder $E(\cdot)$, producing intermediate feature maps:

$$F^z_{\text{ref}} = E(I^z_{\text{ref}}), \quad F^z_{\text{mov}} = E(I^z_{\text{mov}}), \quad \text{for } z \in \{1, k+1, 2k+1, \ldots\}. \tag{1}$$

The features for the skipped slices are then estimated via linear interpolation along the slice axis to form a complete 3D feature volume. When $k = 1$, all slices are encoded and no interpolation is applied. Here, $E(\cdot)$ denotes the encoder part of foundation models (e.g., DINO-v2, SAM), typically based on vision transformers (ViTs). For each encoder, the output feature map $F^z \in \mathbb{R}^{N \times D}$ consists of N patch tokens and, optionally, a global CLS token, each of dimension D.

Dimension Reduction. In the latent space, the feature dimension D is typically high (e.g., 256 for SAM and 384 for DINO-v2), which can introduce noise and lead to significant computational cost during registration. To address this, we apply principal component analysis (PCA) to reduce the feature dimensionality for each patch token. Let $F^z \in \mathbb{R}^{N\times D}$ denote the feature matrix of slice z, where N is the number of patch tokens. We collect features from all sampled slices of both the reference and moving volumes and concatenate them along the patch dimension, resulting in a joint matrix $F_{\text{all}} \in \mathbb{R}^{M\times D}$, where $M = N \times Z \times 2$. PCA is then computed on this combined matrix, and the resulting projection is applied slice-wise to both volumes. Let d be the number of retained principal components (e.g., $d = 12$), the reduced feature maps become:

$$\tilde{F}^z_{\text{ref}} = F^z_{\text{ref}} \cdot W_{\text{PCA}}, \quad \tilde{F}^z_{\text{mov}} = F^z_{\text{mov}} \cdot W_{\text{PCA}}, \quad \tilde{F}^z_{\text{ref}}, \tilde{F}^z_{\text{mov}} \in \mathbb{R}^{N\times d}, \tag{2}$$

Here, $W_{\text{PCA}} \in \mathbb{R}^{D\times d}$ is the projection matrix computed from PCA on F_{all}. This ensures both volumes are embedded into the same reduced feature space.

To prepare for volumetric registration, we reshape $\tilde{F}^z_{\text{ref}}$, $\tilde{F}^z_{\text{mov}}$ back to a 3D spatial format. Specifically, each slice-level feature map $\tilde{F}^z \in \mathbb{R}^{N\times d}$ is first reshaped into a 2D grid of size $w \times h \times d$, where w and h represent the number of patch tokens along the width and height, respectively. Stacking the reshaped slices along the z-axis results in a 4D volume. This intermediate volume is then upsampled to match the original spatial resolution, resulting in a final feature volume of size $W \times H \times Z \times d$, which aligns with the input dimensions and serves as input to the registration step.

Feature-Based Registration. We follow ConvexAdam [23], adopting a two-stage registration strategy that combines global discrete optimization with instance-level refinement to efficiently align the reduced feature volumes.

In the first stage, we employ a coupled convex discrete optimization framework to obtain an initial deformation field based on the extracted feature volumes. A convex energy function is formulated, which combines the feature similarity term with a regularization term to ensure smooth deformations:

$$\mathcal{E}(\phi) = -\text{Sim}\left(\tilde{F}_{\text{ref}}, \tilde{F}_{\text{mov}} \circ \phi\right) + \lambda \|\nabla\phi\|^2 \tag{3}$$

Here, ϕ is the dense displacement field, $\text{Sim}(\cdot,\cdot)$ denotes a similarity metric (e.g., sum of squared differences or local cross-correlation), and λ controls the smoothness constraint. The convex optimization is performed within a discrete search space to efficiently identify the globally optimal deformation. In this work, we set $\lambda = 2$ and MSE for Sim, consistent with DINO-Reg [24].

In the second stage, the deformation field is refined through continuous instance-wise optimization via Adam [12]. Initialized with the displacement field from the first stage, we iteratively update ϕ by minimizing the same energy function in a differentiable manner, allowing for sub-voxel accuracy and capturing fine-scale deformations: $\phi^* = \arg\min_{\phi} \mathcal{E}(\phi)$. This optimization strategy utilizes the robustness of discrete search for global alignment and the flexibility of continuous optimization for local refinement.

2.2 Foundation Models

We use five publicly available foundation models as feature extractors, varying in pre-training strategy and domain (natural vs. medical), as shown in Fig. 2.

DINO-v2. [20]: Pre-trained on natural images using a self-supervised contrastive learning strategy (self-distillation without labels). DINO-v2 is known for producing semantically rich features and has been applied in one prior work for cross-domain medical image registration [24].

SAM. [13]: Trained on large-scale natural image segmentation datasets using a promptable segmentation objective. The model adopts a ViT-based encoder and was first pre-trained with MAE, followed by additional prompt-based segmentation training. SAM has also been widely used for general vision tasks. Although SAM is designed for prompt-based segmentation, we use only its image encoder to automatically generate latent embeddings without human input.

MedSAM. [17]: A domain-specific version of SAM, fine-tuned on various medical image datasets including MRIs under a prompt-based setting to improve performance on medical image segmentation tasks. Notably, as shown in the original MedSAM

Tables 14, the training datasets did not include *any* breast MRI or breast CT images.

SSLSAM. [10]: Fine-tuned from SAM's initial weights on various MRI datasets without labels under the MAE setting. Noticeably, their pre-trained dataset contains one publicly available breast dataset for dynamic pre-contrast MRIs, but not other breast MRI sequences, *e.g.,*, post-contrast [16].

MedCLIP-SAM. [15]: A vision-language foundation model trained on paired medical images and radiology reports. It follows the same training strategy as MedCLIP, but while MedCLIP was only trained on radiographs, MedCLIP-SAM includes additional MRI data during pretraining.

3 Experiments

3.1 Datasets Preparation

Exam Collection. For the MRI-to-MRI registration tasks, the data is collected internally from our institution, consisting of exams performed under the bilateral breast MRI protocol, ranging from the year 2014 to the year 2021. We only include patients who have multiple time-stamped exams and at least one pre-contrast axial view volume available. For the PET-CT to MRI registration task, the data is collected externally from BREAST-DIAGNOSIS collection [4], where we select patients who have both PET-CT and MRI scans acquired within a six-month interval. All exams are first resampled into $1mm \times 1\,mm \times 1\,mm$ isotropic voxel spacing and then padded into $512 \times 512 \times 512$ as volume size. For each task, we randomly selected fifty pairs from the pairs that meet the task criteria, with

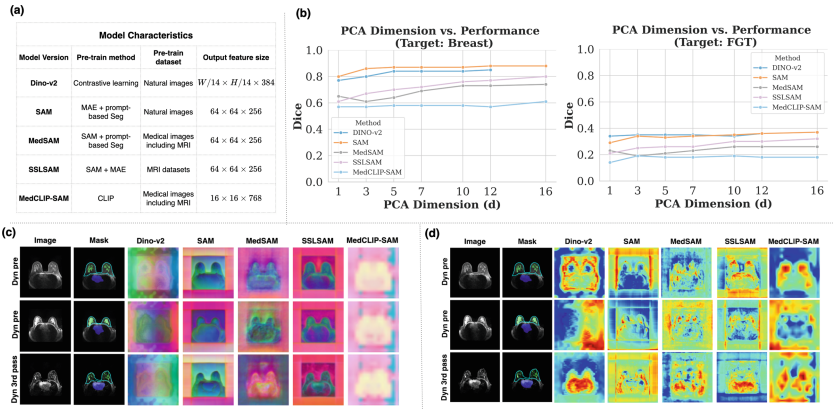


Fig. 2. (a) Encoder characteristics for each VFM (training strategy and dataset domain). (b) Ablation of dimensionality reduction (PCA dimension d) versus registration performance for each encoder. (c) First three channels of the dimension-reduced feature map, visualized as RGB. (d) Top feature channel from raw feature map showing the highest FGT contrast, aiming to visualize whether FGT-related information is captured in the latent features

the task details described in a later paragraph *Tasks*. From these pairs, volumes from patients who had undergone mastectomy were excluded.

Mask Collection. We include three targets for our registration evaluation: breast, FGT, and organ. The first two, breast and FGT, are important targets to register for breast-specific tasks, as accurate alignment of both overall breast shape and internal dense tissue is critical for clinical and research use. For the organ label, we suspect that pre-trained networks may perform relatively well, as this type of structure is commonly seen in abdominal MRI or CT datasets that many of these models were originally exposed to. For breast and FGT mask, we employ the breast segmentation model given by the previous study utilizing 3D V-Net architecture [7,16] followed by the manual correction. The organ and lesion masks are generated using a publicly available breast MRI segmentation model based on nnU-Net, followed by manual correction. The generated masks are visualized in the right-hand side of Fig. 2, in the *Mask* column.

Tasks Design. In this study, we design *four* tasks and one ablation study to comprehensively evaluate the registration performance of multiple models. *Task 1* involves registering breast MRI scans from different dates and years (longitudinal exams) with the same image sequence (all images are dynamic pre-contrast). *Task 2* focuses on registering longitudinal breast MRI exams with different image sequences, where the fixed images are set to the dynamic pre-contrast sequence. The category of moving image sequences is shown in Fig. 3 (b). *Task 3* explores lesion tracking, where we register the moving image with the lesion to the fixed image without the lesion. This task is designed to evaluate whether the registration model attempts to erase or suppress the lesion in the moving image

in order to match the target. For *Task 4*, we register PET-CT to MRI, which is more challenging due to both the cross-modality nature and the significant breast deformation caused by different patient postures (lying on the back versus facing down). We also include the ablation study to disentangle the effects of feature representation and optimization strategy, we replaced ConvexAdam with standard registration methods, SyN and NiftyReg FFD, to register SAM-derived feature volumes in three randomly selected volume pairs from Task 2. The result for all tasks are shown in Table 1.

Evaluation Metrics. *Tasks 1* to *Task 3* are evaluated using the Dice Similarity Coefficient (DSC) of three targets: breast, FGT, and organ. For *Task 3*, we further assess lesion preservation by calculating $V_\varepsilon = abs(\sum M_{les-warped} - \sum M_{les-mov})/\sum M_{les-mov}$, which reflects the change in lesion size during registration. For lesion tracking purposes, the ideal goal is to align the breast and dense tissue while preserving the lesion's physical characteristics, rather than altering its shape or size during the registration process. For *Task 4*, we qualitatively evaluate the registration performance by visualizing the warped CT and overlaying the PET onto the fixed MRI.

3.2 Implementation Details of VFM-Based Registration

For SAM-related models, including SAM, MedSAM, and SSL-SAM, the input slices are resized to 1024×1024. For DINO-v2, the input size is flexible, so we upsample the slices to the largest size that can fit into a single GPU A6000, which is 1792×1792. For MedCLIP-SAM, the input slices are resized to 224×224. All inputs are normalized according to the pretraining settings of each model, using the normalization methods provided in their respective codebases. For the *dimension reduction* step, the size of the extracted feature map for each slice (F^z) varies by model. The PCA dimension d for each encoder is optimized on 9 sampled pairs from *Task 3*. We select the d at which DSC plateaus, balancing performance and computational cost, as higher d increases memory usage and slows registration (Fig. 2). Based on this, we set $d = 16$ for SAM, $d = 10$ for MedSAM, $d = 16$ for SSLSAM, $d = 10$ for DINO-v2, and $d = 16$ for MedCLIP-SAM. The choice of d is based on an ablation from $d = 1$ to 16 (capped at $d = 12$ for DINO-v2 due to GPU limits). For the *registration* step, we follow the hyperparameters selected by DINO-reg [24].

4 Results and Discussion

Results. Table 1 shows that for same-sequence MRI registration (Task 1), foundation models, especially SAM, achieve superior performance in large-structure registration like the breast contour, reaching a DSC of 87.0%, and clearly outperforming the best optimization-based method. For FGT, we observe that SAM and SSLSAM perform better in *Task 1*, while other VFM-based models show a

Table 1. Comparison of Methods on Three Targets Using DSC(%) and Error Size(%)

Method		Breast	FGT	Organ	Lesion
		DSC ↑	DSC ↑	DSC ↑	V_ε ↓
Task 1: Registration between longitudinal scans within the same MRI sequence					
Init.	Raw	49.0 ± 20.6	21.2 ± 16.6	48.3 ± 21.5	-
Opt.	Affine	68.2 ± 21.7	33.0 ± 21.3	60.9 ± 22.8	-
	Syn	80.0 ± 20.0	45.4 ± 24.9	66.7 ± 23.6	-
	NiftyReg	81.6 ± 17.9	53.8 ± 26.5	69.2 ± 24.2	-
	ConvexAdam	80.8 ± 16.6	53.3 ± 23.4	64.7 ± 21.3	-
Found.	DINO-v2	81.3 ± 11.7	47.4 ± 21.8	66.6 ± 19.4	-
	SAM	**87.0 ± 8.3**	53.8 ± 17.6	69.1 ± 18.4	-
	MedSAM	80.9 ± 11.3	47.3 ± 21.6	68.9 ± 18.3	-
	SSLSAM	82.1 ± 13.7	**55.0 ± 22.7**	**72.3 ± 16.9**	-
	MedCLIP-SAM	69.7 ± 13.8	33.4 ± 21.5	59.1 ± 16.1	-
Task 2: Registration between longitudinal scans with different MRI sequencess					
Init.	Raw	44.1 ± 24.4	18.7 ± 13.4	45.3 ± 21.6	-
Opt.	Affine	65.8 ± 22.3	32.0 ± 18.5	61.1 ± 22.7	-
	Syn	76.8 ± 19.1	42.8 ± 21.7	64.4 ± 21.6	-
	NiftyReg	73.9 ± 22.2	44.3 ± 25.0	62.1 ± 22.0	-
	ConvexAdam	73.4 ± 10.3	42.4 ± 31.9	41.2 ± 35.4	-
Found.	DINO-v2	80.4 ± 11.4	44.5 ± 18.9	62.2 ± 18.9	-
	SAM	**85.3 ± 9.0**	**50.2 ± 19.4**	66.4 ± 17.7	-
	MedSAM	76.3 ± 15.1	40.0 ± 19.0	**66.5 ± 18.3**	-
	SSLSAM	81.5 ± 12.1	45.7 ± 19.1	63.6 ± 21.6	-
	MedCLIP-SAM	59.8 ± 18.7	18.5 ± 15.9	50.6 ± 16.3	-
Task 3: Registration from lesion-present cases to lesion-absent cases					
Init.	Raw	49.4 ± 20.3	18.1 ± 14.3	40.4 ± 21.9	0 (ref)
Opt.	Affine	68.0 ± 15.6	31.2 ± 18.2	53.8 ± 24.4	0 (ref)
	Syn	80.5 ± 10.1	41.5 ± 21.5	57.6 ± 24.8	37.7 ± 28.6
	NiftyReg	81.0 ± 9.9	48.4 ± 22.3	57.5 ± 25.5	48.9 ± 30.3
	ConvexAdam	78.3 ± 18.9	**50.3 ± 24.4**	53.3 ± 27.3	60.5 ± 85.5
Found.	DINO-v2	80.5 ± 10.6	40.8 ± 18.7	55.1 ± 22.0	**31.3 ± 23.3**
	SAM	**86.9 ± 8.6**	46.7 ± 21.2	**60.0 ± 23.9**	48.4 ± 32.8
	MedSAM	79.7 ± 12.6	39.3 ± 21.0	59.3 ± 23.7	46.7 ± 32.2
	SSLSAM	85.2 ± 9.9	45.7 ± 21.2	59.1 ± 23.7	60.9 ± 41.2
	MedCLIP-SAM	69.6 ± 17.9	26.6 ± 18.9	49.0 ± 21.4	98.6 ± 150.8
Ablation study: contribution of features vs. optimizer					
	Raw + SyN	71.1 ± 30.4	40.9 ± 18.5	65.7 ± 30.0	-
	Raw + NiftyReg	70.5 ± 20.4	**50.5 ± 19.0**	70.7 ± 6.5	-
	Raw + ConvexAdam	69.4 ± 25.4	46.2 ± 22.7	63.7 ± 11.3	-
	SAM + SyN	58.9 ± 32.7	24.7 ± 11.1	53.7 ± 24.8	-
	SAM + NiftyReg	74.9 ± 26.5	35.7 ± 20.7	69.7 ± 10.7	-
	SAM + ConvexAdam	**85.0 ± 18.8**	47.0 ± 27.9	**73.4 ± 9.2**	-

decreased performance. For organ registration, as we initially hypothesized, models pre-trained on medical-domain data perform better, with SSLSAM achieving the highest performance due to its exposure to similar anatomical structures.

In cross-sequence registration (*Task 2*), SAM achieves a DSC improvement of 8.5% for breast, 5.9% for FGT, and 2% for organ compared to the best optimization-based method, with further visualizations of performance across different sequence subgroups shown in Fig. 3 (b). This suggests that features extracted by foundation models are domain-invariant, making them more robust under shifts in image appearance. In contrast, optimization-based methods rely more heavily on image-level similarity.

Regarding *Task 3* (lesion tracking), DINO-v2 achieves the best performance in preserving lesion size, while MedCLIP-SAM performs poorly. For *Task 4* (PET-CT to MRI registration), the benefits of foundation models become more evident. As shown in Fig. 3 (d), NiftyReg completely fails to align organs between CT and MRI, SAM successfully registers CT to MRI despite significant shape differences, confirming the advantage of foundation models under large domain gaps (e.g., CT to MRI) [24]. This result also suggests potential applications in overlaying PET onto MRI to visualize tumors.

Discussion for pre-train Strategies. Although both SAM and DINO-v2 are pre-trained on natural images, we observe a performance gap here. We attribute this to differences in their training strategies: SAM uses a two-stage pipeline, where the second stage is explicitly trained for pixel-wise segmentation, enabling it to learn fine-grained, spatially localized, and structure-aware representations. DINO-v2, on the other hand, is trained via contrastive learning. While it includes local and global tokens, the patch sizes are relatively large, limiting its ability to capture detailed spatial information. Compared to MedCLIP-SAM, the feature tokens are even coarser, and the model is originally designed for image-text alignment rather than local-information preserving. To further understand these differences, we visualize the feature maps from each model in Fig. 2. For row 3, only SAM and SSLSAM capture some FGT structures, while others fail. Even for SAM/SSLSAM, FGT details is barely visible in the feature map, this limitation is also evident in the qualitative results (Fig. 3 (a)), where models generally register the breast contour well, but substantial FGT misalignment remains. The zero-shot performance of the best VFM-based method still lags significantly behind other training-based approaches, such as GuidedMorph [6].

Discussion for Pre-train Datasets. When comparing models trained on natural images versus those trained on medical datasets (e.g., MedSAM) or pre-contrast breast MRIs (SSLSAM), we observe mixed results. SSLSAM shows improvement in FGT and organ registration in *Task 1*, but performance drops in *Task 2*. We suspect this is due to SSLSAM's pretraining solely on pre-contrast breast MRIs [10], leading to domain overfitting. As a result, while SSLSAM performs reasonably well for pre-contrast MRI alignment (*Task 1*), it performs worse other unseen sequences in *Task 2*. This is also evident in Fig. 3(b), where SSLSAM performs better on sequences similar to its pretraining data but shows noticeably worse performance on others, such as SUB. For MedSAM, we observe substantial performance drops across all tasks for breast and FGT. This likely results from its pretraining on a limited set of medical imaging datasets (approximately 1 million images, compared to 11 million natural images used for SAM),

none of which included breast MRIs or CT, as indicated in the original MedSAM paper. However, it contains substantial data for abdominal MRIs, which makes it capable of capturing organ features effectively. As shown in Fig. 2, its feature maps primarily emphasize organ structures. MedCLIP-SAM fails across all tasks despite being trained on MRIs. Due to the low resolution of its feature maps (Fig. 2), it is far from suitable for registration.

Overall, we do not observe clear benefits from using medical-domain pre-trained models for these breast MRI registration tasks. One possible explanation is that these models, despite being trained on medical data, still use much smaller datasets compared to large-scale natural image pre-training used by models like DINO-v2 or SAM, resulting in less generalizable features. We do not draw definitive conclusions at this stage regarding the effectiveness of medical pre-trained models for zero-shot registration tasks, as new models continue to be proposed [8,25]. Further exploration and discussion are needed to better understand and improve foundation models for this application.

Discussion About Ablation Study. We conducted two groups of ablation studies. The first explores the effect of PCA dimension (d) on registration performance, as shown in Fig. 2. We observed that when d is too low, it may discard important anatomical details necessary for accurate registration, as illustrated by the first three dimensions in Fig. 2 (c). However, as d increases, the performance plateaus and reaches a relatively stable level. Since larger d values significantly increase memory usage and computation time, we find that selecting d in the range of 1016 offers a good trade-off for each encoder. The second study evaluates the contribution of feature maps and optimizers to registration performance. As shown in Table 1, SAM's extracted features outperform raw images, and the combination of SAM features with the ConvexAdam optimizer achieves the best performance among all VFM-based registration methods.

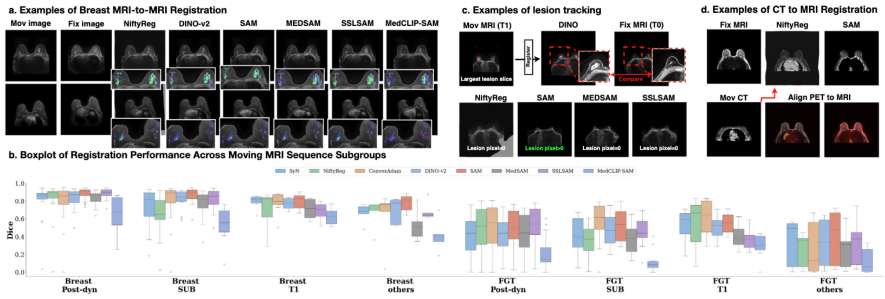


Fig. 3. (a). Visualization of example results for different VFM-based algorithms. FGT from the fixed image slice is shown in blue, and the warped dense tissue is shown in purple. The overlapping regions appear in green; (b). subgroup analysis of the registration performance versus the sequence of the moving image; (c)-(d). Visualization results of selected cases for task 3 and task 4 (Color figure online)

Conclusion. In this study, we explored whether foundation models can be used for zero-shot registration tasks in the context of breast MRI. We find that models pre-trained on natural images, such as SAM, can achieve strong performance for large-structure alignment, particularly for breast contour registration. However, these models still struggle to preserve fine anatomical details, such as FGT, limiting their application in tasks that require accurate tracking of internal structures. We believe this highlights an important direction for future research, which could develop strategies to better preserve fine-grained information within the feature representations of foundation models.

Acknowledgments. The research reported in this publication was supported by the National Institute of Biomedical Imaging and Bioengineering of the National Institutes of Health under the award number R01EB031575. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

1. Alfano, F., et al.: Breast tumor localization by prone to supine landmark driven registration for surgical planning. IEEE Access **10**, 122901–122911 (2022)
2. Avants, B.B., Epstein, C.L., Grossman, M., Gee, J.C.: Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. Med. Image Anal. **12**(1), 26–41 (2008)
3. Balakrishnan, G., Zhao, A., Sabuncu, M.R., Guttag, J., Dalca, A.V.: VoxelMorph: a learning framework for deformable medical image registration. IEEE Trans. Med. Imaging **38**(8), 1788–1800 (2019)
4. Bloch, B.N., Jain, A., Jaffe, C.C.: Breast-Diagnosis. Data Set (2015)
5. Chen, J., Frey, E.C., He, Y., Segars, W.P., Li, Y., Du, Y.: TransMorph: transformer for unsupervised medical image registration. Med. Image Anal. **82**, 102615 (2022)
6. Chen, Y., et al.: GuidedMorph: two-stage deformable registration for breast MRI. arXiv preprint arXiv:2505.13414 (2025)
7. Chen, Y., et al.: Breast density in MRI: an AI-based quantification and relationship to assessment in mammography. arXiv preprint arXiv:2504.15192 (2025)
8. Dong, H., et al.: MRI-Core: a foundation model for magnetic resonance imaging. arXiv preprint arXiv:2506.12186 (2025)
9. Dong, H., Gu, H., Chen, Y., Yang, J., Chen, Y., Mazurowski, M.A.: Segment anything model 2: an application to 2D and 3D medical images. arXiv preprint arXiv:2408.00756 (2024)
10. Gu, H., Dong, H., Yang, J., Mazurowski, M.A.: How to build the best medical image segmentation algorithm using foundation models: a comprehensive empirical study with segment anything model. arXiv preprint arXiv:2404.09957 (2024)
11. Kim, B., Kim, D.H., Park, S.H., Kim, J., Lee, J.G., Ye, J.C.: CycleMorph: cycle consistent unsupervised deformable image registration. Med. Image Anal. **71**, 102036 (2021)
12. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2015), http://arxiv.org/abs/1412.6980

13. Kirillov, A., et al.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4015–4026 (2023)
14. Klein, S., Staring, M., Murphy, K., Viergever, M.A., Pluim, J.P.: Elastix: a toolbox for intensity-based medical image registration. IEEE Trans. Med. Imaging **29**(1), 196–205 (2009)
15. Koleilat, T., Asgariandehkordi, H., Rivaz, H., Xiao, Y.: MedCLIP-SAMv2: towards universal text-driven medical image segmentation (2025). https://arxiv.org/abs/2409.19483
16. Lew, C.O., et al.: A publicly available deep learning model and dataset for segmentation of breast, Fibroglandular tissue, and vessels in breast MRI. Sci. Rep. **14**(1), 5383 (2024)
17. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nat. Commun. **15**(1), 654 (2024)
18. Mazurowski, M.A., Dong, H., Gu, H., Yang, J., Konz, N., Zhang, Y.: Segment anything model for medical image analysis: an experimental study. Med. Image Anal. **89**, 102918 (2023)
19. Mehrabian, H., Richmond, L., Lu, Y., Martel, A.L.: Deformable registration for longitudinal breast MRI screening. J. Digit. Imaging **31**, 718–726 (2018)
20. Oquab, M., et al.: DINOv2: Learning robust visual features without supervision (2024). https://arxiv.org/abs/2304.07193
21. Ringel, M.J., Richey, W.L., Heiselman, J.S., Luo, M., Meszoely, I.M., Miga, M.I.: Supine magnetic resonance image registration for breast surgery: insights on material mechanics. J. Med. Imaging **9**(6), 065001–065001 (2022)
22. Shi, P., Qiu, J., Abaxi, S.M.D., Wei, H., Lo, F.P.W., Yuan, W.: Generalist vision foundation models for medical imaging: A case study of segment anything model on zero-shot medical segmentation. Diagnostics **13**(11), 1947 (2023)
23. Siebert, H., Großbröhmer, C., Hansen, L., Heinrich, M.P.: ConvexAdam: self-configuring dual-optimisation-based 3D multitask medical image registration. IEEE Trans. Med. Imaging **44**(2), 738–748 (2024)
24. Song, X., Xu, X., Yan, P.: DINO-Reg: general purpose image encoder for training-free multi-modal deformable medical image registration. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 608–617. Springer (2024)
25. Sun, Y., Wang, L., Li, G., Lin, W., Wang, L.: A foundation model for enhancing magnetic resonance images and downstream segmentation, registration and diagnostic tasks. Nat. Biomed. Eng. **9**(4), 521–538 (2025)



Towards Breast Cancer Recurrence Prediction Using Transformer-Based Learning from Global–Local Radiomics and Clinical Data

Adnan Khalid[1(✉)], Muhammad Mursil[1], Fabrice Meriaudeau[2], Ibrahima Faye[3], Alain Lalande[2], Domenec Puig[1], and Hatem A. Rashwan[1]

[1] DEIM, IRCV, Universitat Rovira i Virgili, Tarragona, Spain
{adnan.khalid,hatem.abdellatif}@urv.cat
[2] ICMUB Laboratory, UMR CNRS, Université de Bourgogne, Dijon, France
[3] CISIR, Universiti Teknologi PETRONAS, Perak, Malaysia

Abstract. Accurate prediction of breast cancer recurrence remains a major challenge in precision oncology, primarily due to the biological heterogeneity of tumor subtypes and the limitations of current genomic assays. In this work, we introduce a dual-stream, multimodal framework that combines radiomic features extracted from both whole-breast parenchymal tissue and segmented tumor regions with clinicopathological variables. Leveraging a fully automated pipeline for segmentation and feature extraction, we compute enhancement, symmetry, and texture descriptors from DCE-MRI volumes, alongside curated clinical data. A transformer-based model (TabNet) is employed for recurrence risk prediction, demonstrating superior discriminative performance (AUC of 0.8193, F1 of 0.7143) compared to conventional machine learning approaches. Feature attribution via SHAP confirms the prognostic relevance of bilateral symmetry, tumor texture, and hormone receptor status. These findings underscore the potential of interpretable, multimodal deep learning for non-invasive, scalable, and individualized recurrence risk stratification (https://github.com/adnankhalid7454/BC_Recurrence_Pred.).

Keywords: Breast Cancer Recurrence · Interpretable AI · Radiomics · DCE-MRI · Breast Segmentation

1 Introduction

Breast cancer (BC) remains the most frequently diagnosed malignancy among women worldwide, with over 2.3 million new cases reported annually. Despite significant progress in early detection and therapeutic interventions, BC continues to be a leading cause of cancer-related mortality [15]. A key challenge

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
T. Zhang et al. (Eds.): Deep-Breath 2025, LNCS 16142, pp. 113–122, 2026.
https://doi.org/10.1007/978-3-032-05559-0_12

in its clinical management is the marked biological heterogeneity of the disease, particularly among molecular subtypes such as human epidermal growth factor receptor 2-positive (HER2+) and triple-negative breast cancer (TNBC). These subtypes are characterized by aggressive clinical behavior, early recurrence, and diverse treatment responses [2]. Although systemic chemotherapy is standard for these subtypes—especially TNBC—its efficacy varies considerably across patients, and recurrence remains common, motivating efforts to improve risk stratification [3,11].

Accurate prediction of BC recurrence is essential for optimizing adjuvant therapy decisions, reducing overtreatment, and enabling personalized follow-up strategies. Existing prognostic tools, such as multigene expression assays (e.g., Oncotype DX and MammaPrint), provide valuable insights but are limited by high cost, invasive sampling requirements, and limited availability in specific healthcare settings [4,14].

Dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI) has emerged as a powerful modality that not only characterizes tumor morphology but also captures functional attributes, including vascularity, enhancement kinetics, and background parenchymal enhancement (BPE), all of which are increasingly associated with tumor biology and prognosis [7,17] . Recent studies have explored radiomic signatures derived from DCE-MRI that correlate with recurrence-related genomic assays. For example, Li et al. [8] demonstrated that features capturing tumor size, enhancement texture, and margin morphology were significantly associated with recurrence scores from Oncotype DX, PAM50, and MammaPrint, achieving area under the curve (AUC) values of up to 0.88. Saha et al. [13] conducted a large-scale radiogenomic analysis involving 922 patients and 529 imaging features, revealing moderate correlations between MRI-derived phenotypes and molecular subtype, ER/PR status, and the Ki-67 proliferation index (AUCs of approximately 0.65 to 0.70). Romeo et al. [12] used handcrafted radiomics features from segmented tumors in DCE-MRI and applied logistic regression to predict Oncotype DX scores, reporting a test-set accuracy of 63% with an AUC of 0.66. Similarly, Nam et al. [10] developed a radiomic score using LASSO-selected features from 3T DCE-MRI, achieving an AUC of 0.76 for distinguishing low and non-low Oncotype recurrence risk in ER+ patients.

While the previous studies have shown promising associations between radiomic features from DCE-MRI and recurrence-related genomic signatures, several key challenges remain unresolved. First, most existing approaches focus exclusively on tumor-centric features, often neglecting the prognostic value of global parenchymal tissue characteristics such as bilateral symmetry and background enhancement, which reflect the underlying tumor microenvironment. Second, many models rely on handcrafted feature selection or classical machine learning (ML) classifiers that may fail to capture complex, nonlinear interactions between imaging phenotypes and clinical variables. Third, interpretability—an essential factor for clinical adoption—remains limited in deep learning applica-

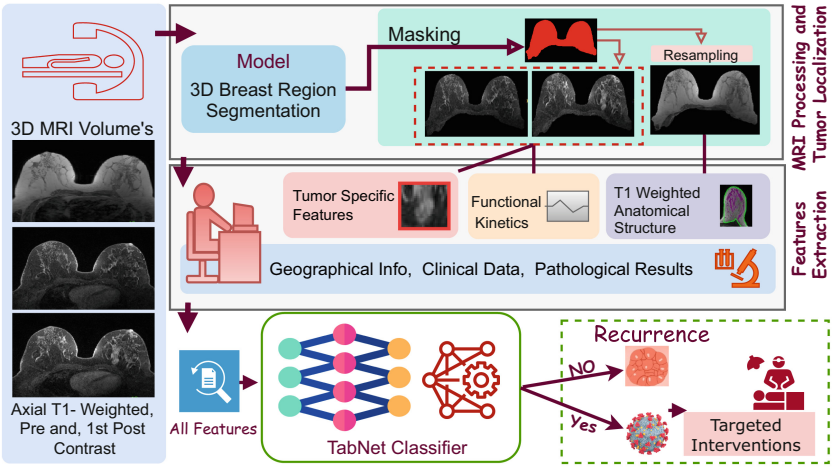


Fig. 1. Overview of the proposed dual-stream framework for BC recurrence prediction. The pipeline includes automated preprocessing and segmentation of DCE-MRI volumes to extract radiomic features from both whole-breast and tumor-specific regions. These features are fused with clinical and pathological data to build a multimodal representation, enabling recurrence classification using the TabNet model.

tions, particularly when applied to tabular data that involves mixed data types and heterogeneous sources.

To address these challenges, this paper presents a dual-stream feature extraction framework that integrates both tumor-specific radiomic features, obtained through automated lesion segmentation, and global parenchymal descriptors extracted from the entire breast volume via a separate segmentation pipeline. This design enables the simultaneous modeling of local tumor morphology and enhancement dynamics, as well as global tissue characteristics such as bilateral symmetry and background enhancement. By combining these imaging features with curated clinicopathological variables, we construct a unified, multimodal representation of recurrence risk. We evaluate multiple machine learning models, including Random Forest, XGBoost, and CatBoost, and benchmark them against TabNet—a transformer-based architecture optimized for tabular data—to assess predictive performance and model interpretability. The proposed approach offers a scalable, interpretable, and fully automated solution for individualized breast cancer recurrence prediction.

2 Materials and Methods

2.1 Dataset

This study utilized the publicly available breast DCE-MRI dataset published by Saha et al. [13], comprising 922 subjects diagnosed with invasive breast can-

cer. Each patient underwent pre-operative DCE-MRI and had associated clinicopathological data and recurrence outcomes recorded at follow-up. Imaging was performed using both 1.5T and 3T MRI scanners, following a standardized dynamic contrast-enhanced protocol: a pre-contrast T1-weighted sequence followed by four post-contrast acquisitions after intravenous gadolinium administration. A cohort of 567 subjects was selected based on the completeness of clinical variables, definitive recurrence status. From the selected cohorts, 514 were assigned to the training set and 53 to the test set (90/10 split). While the training set reflected the real-world class imbalance (64 recurrences, 450 non-recurrences), the test set was enriched with recurrence cases (15 recurrences, 38 non-recurrences) to enhance clinical variability and assess performance on high-risk subgroups. This sampling is to evaluate model generalization on diverse and high-risk recurrence profiles, prioritizing clinical relevance over strict stratified proportions.

2.2 BC Recurrence Framework

The proposed dual-stream framework consists of five main stages, as illustrated in Fig. 1: MRI preprocessing, segmentation, radiomic feature extraction, clinical data integration, and classification.

1) MRI Preprocessing and Normalization: Each DCE-MRI scan was preprocessed to ensure spatial and intensity consistency across subjects. Temporal alignment was performed across the dynamic series to account for motion or misregistration between pre- and post-contrast acquisitions. All volumes were resampled to a standardized voxel size to enable uniform feature extraction. Intensity normalization was applied to reduce scanner-related variability. The first post-contrast volume—typically capturing peak enhancement—was selected as input for tumor segmentation in subsequent steps.

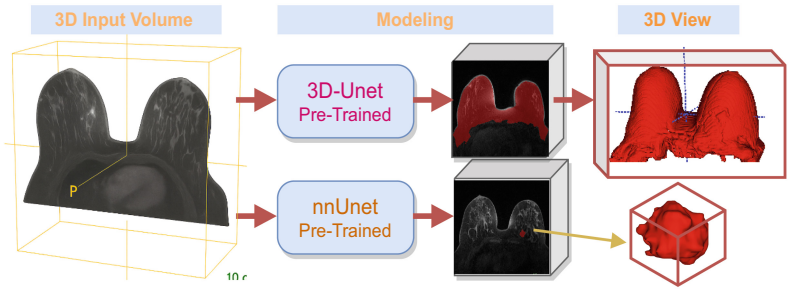


Fig. 2. Visualizations of segmented target regions obtained from trained models.

2) Automated Segmentation: In this work, two fully automated segmentation frameworks were implemented to support downstream analysis and feature extraction. For breast parenchyma segmentation, a fully convolutional neural network—adapted from the architecture proposed by Zhang et al. [16] was

utilized to accurately delineate the breast boundary and fibroglandular tissue (FGT), enabling bilateral symmetry assessment. For tumor segmentation, a pre-trained model introduced by Garrucho et al. [6], trained on a heterogeneous set of annotated DCE-MRI scans, was applied to the first post-contrast volume. The model performed voxel-wise classification to identify and segment-enhancing tumor regions (Fig. 2).

3) Radiomic Feature Extraction: From the segmented breast and tumor masks, a total of 28 quantitative radiomic features were extracted from both the whole-breast and tumor-specific regions. These features, summarized in Table 1, encompass enhancement dynamics [9], parenchymal symmetry, and tumor texture [5] (Energy, Entropy, and Homogeneity), which were quantitatively extracted from the segmented tumor regions using the PyRadiomics library. Collectively, these features are hypothesized to capture biologically relevant phenotypes associated with recurrence risk.

Table 1. Quantitative features from whole-breast (E_{top}: enhancement statistics; Density; Symmetry) and tumor-specific regions (T_{Volume}: tumor volume; Texture features).

Feature Category	Mathematical Definition / Extracted Features
Enhancement Statistics (Whole Breast)	Compute enhancement map: $E = \frac{I_{\text{post}} - I_{\text{pre}}}{I_{\text{pre}}}$ Select top 2% enhanced voxels within breast mask M_B: $E_{\text{top}} = \text{Top}_{2\%}(E[M_B])$ Features: Mean (μ), Median, Max, Std (σ), Skewness (γ_1), Kurtosis (γ_2)
Density and Symmetry (Whole Breast)	Compute breast density using Otsu threshold (T_{Otsu}): Density $= \frac{\sum \mathbf{1}_{I_{\text{t1}} > T_{\text{Otsu}}}}{\lvert M_B \rvert}$ Calculate symmetry from mirrored breast halves (L, R): SymmetryIndex $= 1 - \frac{\sum \lvert L - \text{flip}(R) \rvert}{\sum \lvert L \rvert + \lvert R \rvert + \epsilon}$
Texture Features (Tumor Region)	Tumor Volume: $T_{\text{Volume}} = \sum \mathbf{1}_{E > T_{98\%}} \cdot M_T$ (voxels above 98th percentile enhancement within tumor mask M_T) Radiomic features (Contrast, Energy, Entropy, Homogeneity)

4) Clinical Data Integration: Key clinical and pathological variables—including hormone receptor status ER, PR, HER2 expression, tumor grade and stage, and molecular subtype—were combined with radiomic features to construct a comprehensive multimodal feature set for each patient. Integrating clinical data with imaging-derived features enables the model to capture both anatomical and molecular determinants of breast cancer recurrence. While radiomic features quantify tumor morphology and enhancement patterns, clinical and pathological variables provide complementary biological context—such

as hormone responsiveness and tumor aggressiveness—that may not be directly observable in imaging. This fusion of heterogeneous data sources allows the model to learn more nuanced, patient-specific recurrence patterns, ultimately improving prediction accuracy and clinical relevance.

5) Breast Cancer Recurrence Classification: To classify patients into recurrence and non-recurrence groups, we employed TabNet [1], a transformer-based deep learning model specifically designed for structured tabular data. TabNet leverages sequential feature selection with sparse attention mechanisms, enabling it to handle high-dimensional, multimodal input while maintaining interpretability. Its architecture is particularly well-suited for clinical settings, as it learns transparent decision pathways across heterogeneous data modalities. All extracted features—including radiomic and clinical variables—were concatenated into a unified feature vector.

Prior to model training, we applied univariate feature selection using the `SelectKBest` method with ANOVA F-statistics, identifying 11 features most strongly associated with recurrence. Feature values were normalized using a Min-Max scaler to ensure comparability across modalities. To address class imbalance in the training data, we applied the Synthetic Minority Over-sampling Technique (SMOTE), while ensuring that the test set remained untouched by any augmentation (Fig. 3).

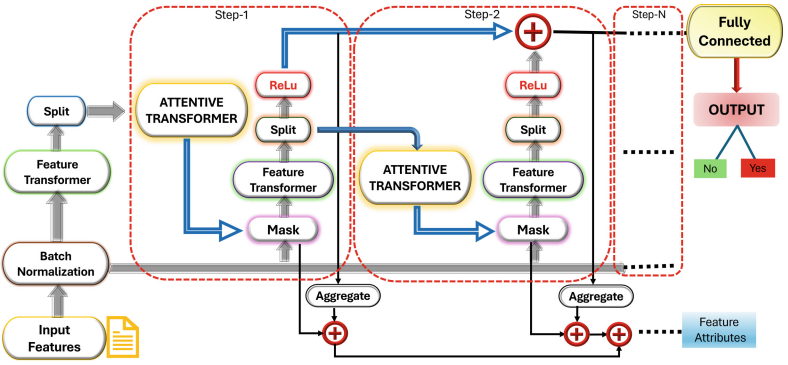


Fig. 3. TabNet architecture utilizes attention masks and feature transformers to learn sparse, interpretable decision paths across multimodal features.

For comparative evaluation, we trained three additional ensemble-based classifiers: Random Forest, XGBoost, and CatBoost. Hyperparameter tuning was performed using grid search with 5-fold cross-validation. Model performance was assessed using accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC). To ensure model transparency, we applied SHAP (SHapley Additive exPlanations) to quantify each feature's contribution to the final predictions. For tree-based models, feature importance was assessed

using the Gini impurity index, while permutation importance was used for TabNet. This multimodal classification pipeline combines tumor-localized phenotypes, global parenchymal patterns, and clinical data, enabling interpretable and robust prediction of BC recurrence.

3 Results and Discussion

Classification Performance: We evaluated four models—TabNet, CatBoost, XGBoost, and Random Forest—on the integrated multimodal feature set comprising radiomic and clinical variables. TabNet demonstrated the most favorable overall performance, achieving an ROC AUC of 0.82 and the highest F1-score (0.71), reflecting the best balance between recall (0.67) and precision (0.77). Although XGBoost yielded the highest ROC AUC (0.83), it exhibited notably low sensitivity (recall = 0.20), limiting its clinical usefulness in accurately detecting recurrence cases. CatBoost and Random Forest showed moderate overall performances with low recall (0.27) and F1-scores (0.40), underscoring their inadequate sensitivity to minority-class predictions (Table 2).

Table 2. Performance comparison of different classifiers on BC recurrence prediction.

Model	Accuracy	Precision	Recall	F1-Score	ROC AUC
Random Forest	0.774	0.800	0.267	0.400	0.756
XGBoost	0.736	0.600	0.200	0.300	0.839
CatBoost	0.774	0.800	0.267	0.400	0.756
TabNet	**0.849**	**0.769**	**0.667**	**0.714**	**0.819**

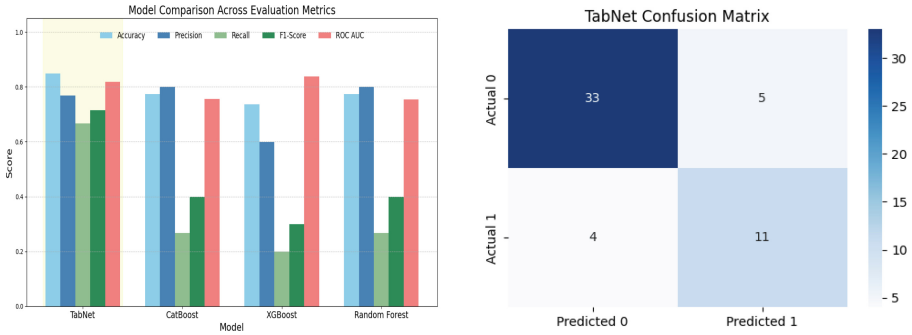


Fig. 4. Bar chart of model performance metrics and confusion matrix of TabNet predictions on the test cohort for BC recurrence classification.

The confusion matrix analysis revealed that TabNet correctly classified 33 out of 38 non-recurrence cases (specificity ≈ 86.8%) and 11 out of 15 recurrence

cases (sensitivity ≈ 73.3%). This strong sensitivity is particularly important in clinical settings, where false negatives may result in missed opportunities for early intervention (Fig. 4). TabNet's superior performance is attributed to its sequential attention mechanism, which efficiently focuses on the most informative clinical and radiomic attributes. Its sparse attention approach effectively manages high-dimensional, heterogeneous datasets, reducing the risk of overfitting.

Ablation Study on Feature Selection: To assess the benefit of pre-filtering features, we compared TabNet's performance with and without applying SelectKBest (ANOVA F-test). Without feature selection (using all extracted features), TabNet achieved an ROC AUC of 0.64. Incorporating SelectKBest to select the top 11 features (empirically determined via nested cross-validation) significantly improved TabNet's ROC AUC to 0.819, confirming the value of preliminary feature reduction in enhancing predictive accuracy.

Segmentation Impact Analysis. To evaluate the influence of segmentation quality, we applied morphological perturbations to tumor masks. This led to a 6.4% drop in classification AUC, highlighting the model's sensitivity to segmentation errors and motivating future use of uncertainty-aware segmentation for improved robustness.

Interpretability and Feature Attribution: To evaluate model transparency and highlight the role of clinical variables, we analyzed feature importance using both built-in TabNet interpretability and SHAP. As shown in Fig. 5, both imaging-derived and clinical features contributed significantly to the recurrence prediction task. Energy Tumor, PR status, and Tumor Grade M emerged as the top three contributors, followed closely by SymmetryIndex and Entropy Tumor. These results highlight the relevance of both texture heterogeneity and bilateral parenchymal asymmetry as imaging biomarkers, alongside key clinicopathological indicators.

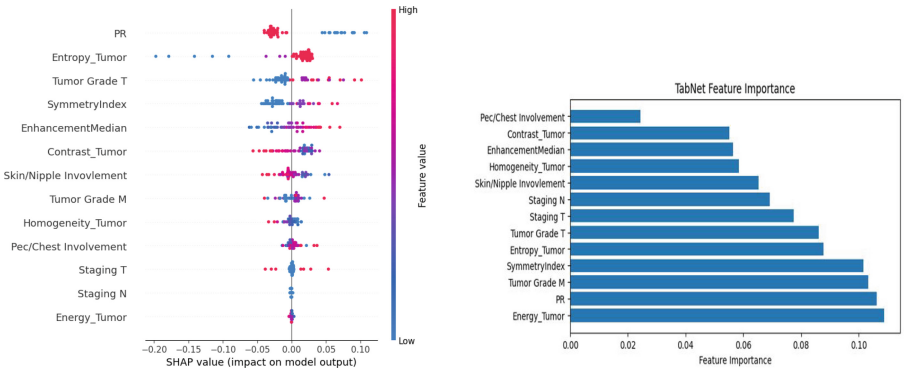


Fig. 5. SHAP analysis and feature importance from the TabNet model, illustrating the relative contributions of radiomic and clinical features to BC recurrence prediction.

Additionally, we used SHAP to quantify the contribution of each feature to the predicted recurrence risk as shown in Fig. 5. The SHAP summary plot revealed that both clinical and imaging-derived features had a strong influence on the model's output. Specifically, higher values of Energy Tumor and Entropy Tumor—representing texture heterogeneity within the tumor—were positively associated with increased recurrence risk. Similarly, elevated SymmetryIndex values, which indicate greater asymmetry between bilateral breast parenchyma, were also predictive of recurrence, likely reflecting tumor-induced tissue distortion. Among clinical features, positive PR status and higher Tumor Grade M were influential, aligning with known prognostic factors. In contrast, lower values of EnhancementMedian and early-stage disease (Staging T) were associated with non-recurrence predictions. These findings confirm TabNet's ability to integrate imaging phenotypes with clinical biomarkers, producing individualized, biologically relevant predictions. Its balanced sensitivity, specificity, and interpretability demonstrate its strong potential for clinical settings, highlighting the value of hybrid models in capturing complex recurrence risks.

4 Conclusion

We presented a dual-stream, multimodal learning framework for BC recurrence prediction that integrates quantitative radiomic features from both whole-breast parenchymal tissue and tumor-specific regions with curated clinicopathological variables. The proposed pipeline employs TabNet, a transformer-based model for tabular data, which consistently outperforms conventional ensemble methods. Notably, it achieved a favorable balance between sensitivity and specificity—addressing a major challenge in recurrence detection—while maintaining interpretability through attention-based feature attribution. Among the most informative predictors were bilateral parenchymal asymmetry, tumor texture heterogeneity, hormone receptor status, and tumor grade. These results underscore the clinical utility of integrating anatomical and molecular-level information for individualized risk stratification. The framework offers a scalable and interpretable solution suitable for clinical decision support. Future work will involve validating the framework on larger, external datasets to support broader generalizability and real-world applicability.

Acknowledgments. This work was supported by the Bosomshield Project, a grant from Marie Sklodowska-Curie Doctoral Networks Actions (HORIZON-MSCA-2021-DN-01-01;101073222).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Arik, S.Ö., Pfister, T.: TabNet: attentive interpretable tabular learning. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 6679–6687 (2021)
2. Bianchini, G., Balko, J.M., Mayer, I.A., Sanders, M.E., Gianni, L.: Triple-negative breast cancer: challenges and opportunities of a heterogeneous disease. Nat. Rev. Clin. Oncol. **13**(11), 674–690 (2016)
3. Dent, R., et al.: Triple-negative breast cancer: clinical features and patterns of recurrence. Clin. Cancer Res. **13**(15), 4429–4434 (2007)
4. Dowsett, M., et al.: Prediction of risk of distant recurrence using the 21-gene recurrence score in node-negative and node-positive postmenopausal patients with breast cancer treated with anastrozole or tamoxifen: a transatac study. J. Clin. Oncol. **28**(11), 1829–1834 (2010)
5. Drukker, K., Li, H., Antropova, N., Edwards, A., Papaioannou, J., Giger, M.L.: Most-enhancing tumor volume by MRI radiomics predicts recurrence-free survival "early on" in neoadjuvant treatment of breast cancer. Cancer Imaging **18**, 1–9 (2018)
6. Garrucho, L., et al.: A large-scale multicenter breast cancer DCE-MRI benchmark dataset with expert segmentations. Sci. Data **12**(1), 453 (2025)
7. Leithner, D., et al.: Clinical role of breast MRI now and going forward. Clin. Radiol. **73**(8), 700–714 (2018)
8. Li, H., et al.: MR imaging radiomics signatures for predicting the risk of breast cancer recurrence as given by research versions of Mammaprint, oncotype DX, and PAM50 gene assays. Radiology **281**(2), 382–391 (2016)
9. Macura, K.J., Ouwerkerk, R., Jacobs, M.A., Bluemke, D.A.: Patterns of enhancement on breast MR images: interpretation and imaging pitfalls. Radiographics **26**(6), 1719–1734 (2006)
10. Nam, K.J., et al.: Radiomics signature on 3T dynamic contrast-enhanced magnetic resonance imaging for estrogen receptor-positive invasive breast cancers: Preliminary results for correlation with oncotype dx recurrence scores. Medicine **98**(23), e15871 (2019)
11. O'Shaughnessy, J.: Extending survival with chemotherapy in metastatic breast cancer. Oncologist **10**(S3), 20–29 (2005)
12. Romeo, V., et al.: MRI radiomics and machine learning for the prediction of oncotype DX recurrence score in invasive breast cancer. Cancers **15**(6), 1840 (2023)
13. Saha, A., et al.: A machine learning approach to radiogenomics of breast cancer: a study of 922 subjects and 529 DCE-MRI features. Br. J. Cancer **119**(4), 508–516 (2018)
14. Sparano, J.A., et al.: Adjuvant chemotherapy guided by a 21-gene expression assay in breast cancer. N. Engl. J. Med. **379**(2), 111–121 (2018)
15. Sung, H., et al.: Global cancer statistics 2020: globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: Cancer J. Clinic. **71**(3), 209–249 (2021)
16. Zhang, J., Saha, A., Zhu, Z., Mazurowski, M.A.: Breast tumor segmentation in DCE-MRI using fully convolutional networks with an application in radiogenomics. In: Medical Imaging 2018: Computer-Aided Diagnosis, vol. 10575, pp. 192–196. SPIE (2018)
17. Zhang, T., et al.: Radiomics and artificial intelligence in breast imaging: a survey. Artif. Intell. Rev. **56**(Suppl 1), 857–892 (2023)

Computer-Assisted Surgical Planning for DIEP Flap Breast Reconstruction Surgery

Jakub Ceranka[1,2(✉)], Diego Lamtenzan[1], Pieter Boonen[1,2,3], Ayush Kapila[4], Carola Brussaard[3], Moustapha Hamdi[4], and Jef Vandemeulebroucke[1,2,3]

[1] Department of Electronics and Informatics (ETRO.RDI), Vrije Universiteit Brussel (VUB), Pleinlaan 2, 1050 Brussels, Belgium
jakub.ceranka@vub.be
[2] imec, Kapeldreef 75, 3001 Leuven, Belgium
[3] Department of Radiology, Universitair Ziekenhuis Brussel (UZ Brussel), Vrije Universiteit Brussel (VUB), Laarbeeklaan 101, 1090 Brussels, Belgium
[4] Department of Plastic, Reconstructive and Aesthetic Surgery, Universitair Ziekenhuis Brussel (UZ Brussel), Vrije Universiteit Brussel (VUB), Laarbeeklaan 101, 1090 Brussels, Belgium

Abstract. The deep inferior epigastric artery perforator (DIEP) flap is the gold standard for autologous breast reconstruction following mastectomy, and involving transplantation of abdominal skin and fat while preserving muscle integrity. A critical yet challenging aspect of surgical planning is the accurate identification of perforator vessels - small arteries supplying blood to the transplanted tissue slab. Manual perforator selection from computed tomography angiography (CTA) scans is time-consuming, prone to error, and frequently resulting in intra-operative adjustments, prolonged surgery and increased risk of complications. We propose a fully-automated pre-operative planning framework designed to reliably map perforator vessels in CTA images. The system consists of a depth-aware annotation strategy that corrects spatial distortions associated with maximum intensity projection imaging; a patient-specific, anatomy-aware region of interest selection; and a deep-learning perforator segmentation pipeline leveraging a Swin UNETR architecture, trained with custom continuity-aware loss function and region-specific patch sampling. The method was validated on clinical CTA scans from 55 patients, using expert-guided annotations of the perforator vessels. The proposed method achieved a perforator segmentation median Dice similarity coefficient of 0.60 and a 95th-percentile Hausdorff distance of less than 18 mm, significantly improving vessel segmentation accuracy in fat regions over baseline Swin UNETR. These results demonstrate the clinical feasibility of automated segmentation, which has the potential to standardise surgical planning, enhance perforator selection precision, and ultimately improve patient outcomes in DIEP flap breast reconstruction.

J. Ceranka and D. Lamtenzan—are shared first author.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
T. Zhang et al. (Eds.): Deep-Breath 2025, LNCS 16142, pp. 123–133, 2026.
https://doi.org/10.1007/978-3-032-05559-0_13

Keywords: DIEP flap surgery · perforator segmentation · computer-assisted surgical planning · deep learning · breast cancer

1 Introduction

Breast cancer remains the leading malignancy in women, with more than 2.3 million diagnoses recorded worldwide in 2022 [12]. Up to 40% of patients still require total breast mastectomy, and one-third of these women cannot accept an implant [8].

The deep inferior epigastric artery perforator (DIEP) flap is an autologous microsurgical technique used for breast reconstruction following total mastectomy, typically performed as a standard procedure for treating or preventing breast cancer. This approach involves transferring skin and subcutaneous fat from the lower abdomen while preserving the rectus abdominis muscle. The DIEP flap facilitates restoration of breast contour, maintains abdominal wall integrity, and offers superior long-term symmetry compared to implant-based reconstruction. It is therefore a current gold standard among surgical breast reconstruction techniques [16]. The demand for the procedure is rising [11], as DIEP flap now accounts for more than 60% of all autologous reconstructions, consuming large blocks of operating-room time, requiring specialised imaging and surgical training, and substantial healthcare resources [3,9]. The amount of tissue that can be safely harvested is typically assessed intra-operatively, relying heavily on surgeon's expertise and complex pre-operative radiological planning.

High-resolution computed tomographic angiography (CTA) is the current imaging standard for pre-operative DIEP flap planning. A single, breath-held acquisition produces sub-millimetre voxels and allows for the visualisation of vessels from 1 mm diameter or greater [17]. CTA enables localization of the DIEP fascial exit points relative to the umbilicus, which are then transposed onto the abdomen using an umbilicus-centered coordinate grid [18].

The surgical planning based on manual slice-by-slice CTA review faces several challenges. A thin-slab maximum intensity projection (MIP) [4] is commonly calculated to increase the perforator contrast. MIPs do however spatially stretch and misplace submillimetre vessels, which appear extended along the direction of MIP mapping but vanish in another. That causes problems with image interpretation and extraction of vessel parameters. The radiologist must manually trace every candidate perforator, record its diameter and intramuscular course, and summarise the findings in a structured report for the surgical team. The workflow is slow (reading times typically exceed 45 min per patient [9]) and remains highly subjective. Moreover, the inter-reader agreement on the perforator selection is low, and surgeons abandon the radiologically selected vessel in roughly one-third of harvests [2]. Such mis-selections prolong dissection, raise fat-necrosis rates, and drive costly revisions [13]. DIEP flap surgery planning could therefore

benefit from automation, reducing the manual burden and planning time, and eliminating image reading variability by providing standardised reporting.

To date, only several studies have addressed aspects of automated planning for DIEP flap surgery, with a primary focus on perforator detection. Lange *et al.* [14] proposed a method in a form of an Osirix image viewer plugin allowing the user to select multiple points on the perforator tree for automated calculation of vessel characteristics (i.e. radius, branching). Araújo *et al.* [1] combined vesselness filtering with minimum-cost path algorithms to semi-automatically track intramuscular and subcutaneous vessel segments, thus reducing reader time. Saxena *et al.* [20] proposed a concept of a computer-aided diagnosis (CAD) system for DIEP perforator tree selection, based on the transfer-learning of the DeepVesselNet [22] and using U-Net [19] and a fully convolutional network [15], however, the model was only used on synthetic data [22] and never applied to DIEP flap CTA imaging and planning. All mentioned methods either rely on manual input, do not provide volumetric information of the vascular tree (i.e. segmentation) or fail to comprehensively address all stages of the surgical planning workflow.

In this work, we propose the first, fully-automated segmentation framework for DIEP flap surgical planning. Our contribution is three-fold: (a) a method addressing commonly existing spatial deformation limitations to perforator annotation from MIP imaging, (b) a patient anatomy-aware cropping method for region of interest selection, (c) a fully-automated segmentation framework for perforator vascular tree mapping from CTA.

2 Materials and Methods

2.1 Dataset

CTA. The dataset consists of 55 fully anonymised, retrospectively collected arterial-phase abdominal CTA from women assessed for DIEP flap breast reconstruction. Images were acquired on a 64-row GE *Discovery CT750 HD* as a part of a routine pre-operative imaging protocol at the Universitair Ziekenhuis Brussel (UZ Brussel). Ethical board approval and written informed consent were obtained prior to data release. The following imaging parameters were used: 120 kVp tube voltage, 523 ± 11 mA tube current, 0.40 s rotation time, pitch 0.98, helical mode. Volumes were reconstructed at 0.63 mm slice thickness with an in-plane pixel size of approximately 0.80 mm and a 360 mm field of view.

All volumes were re-oriented to RAS^+ and resampled to an isotropic $1 \times 1 \times 1$ mm, required for MIP back-projection. To improve perforator contrast, 10 mm slab-MIP projections were generated along the axial and coronal planes of CTA image, prior to the manual perforator annotation process. Voxel intensities were standardised with per-volume standard score (z-score).

Image Annotation. Full perforator vessel trees were independently annotated in the coronal ($n = 55$) and axial ($n = 55$) MIP images (Slicer v5.2.2 [7]) by a trained medical imaging researcher and confirmed by a radiologist with more than 20 years of experience in DIEP flap CTA pre-operative planning.

2.2 CAD Pipeline

The system consist of: (1) a depth-aware annotation strategy that corrects spatial distortions from slab-MIP imaging; (2) a patient-specific, anatomy-aware region of interest (ROI) selection; and (3) a Swin UNETR-based perforator segmentation trained with a continuity-aware loss and region-specific patch sampling (see Fig. 1).

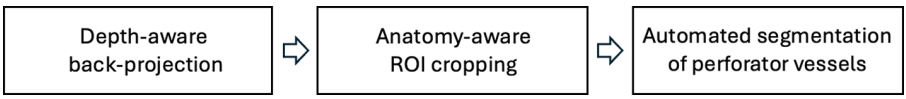


Fig. 1. Overview of the automated three-stage CAD pipeline

Depth-Aware Back-Projection for the Annotation of MIP Images. To correct the spatial stretching and misplacement of submillimetre vessels inherent to thin-slab MIPs and inspired by the depth reverse projection framework of Dima *et al.* [6], we record each MIP voxel's slice index during projection to generate a depth-index volume that maps 3D MIP coordinates back into the 3D CTA space. Axial and coronal annotations in MIP space are back-projected into the original CTA isotropic volume via their depth indices. Reconstructed annotations (coronal and axial) are fused across views to eliminate orientation-dependent discontinuities, and preserve true vessel geometry, improving annotation quality without introducing MIP spacing artifacts. A final manual refinement using morphological operations produces accurate binary masks for training and evaluation (see Fig. 2).

Anatomy-Aware ROI Cropping. To limit computation to the anterior abdominal wall and focus on the region where the DIEP are located, each CTA volume is passed once through the pre-trained *TotalSegmentator* [23] bone CT segmentation model. Acquired bone segmentations of the pelvic and sternum allow for the automated anatomical landmark extraction: the lateral peaks of both iliac crests, the pubic symphysis midpoint, and the lowest sternum voxel (xiphoid). The cranio-caudal ROI is bounded inferiorly at 30% of the vertical distance from the pubic symphysis to the average iliac crest level, and superiorly at 60% of the distance from the pubic symphysis to the xiphoid landmark; the left-right span is bounded by the crests; the anterior margin reaches the skin surface plus a 10-voxel pad obtained by ray-casting anteriorly (see Fig. 3). Based on these landmarks and proportional bounds, a patient-specific ROI is generated that reliably captures the DIEP territory while excluding non-relevant anatomy.

Automated Segmentation of Perforator Vessels.

Baseline Method. Our reference model is the 3D *Swin UNETR* [10]. A four-stage Swin-Transformer encoder (depths $[2, 2, 2, 2]$, heads $[3, 6, 12, 24]$, hidden size 48)

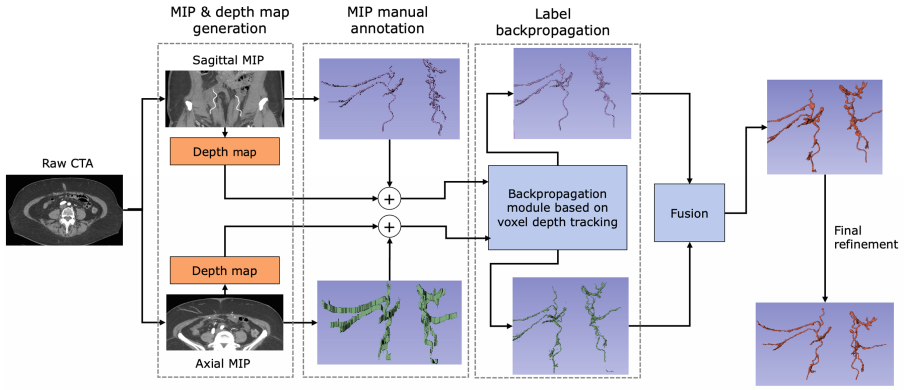


Fig. 2. A schematic overview of the depth-aware MIP label back-projection method. The process begins with a CTA image from which two MIP views are generated. These views are then manually annotated to identify relevant structures. Simultaneously, the depth value of each annotated voxel is recorded. This depth information enables precise back-projection of annotations from the MIP views onto the original 3D CTA volume, effectively restoring the true spatial location of each label without introducing distortions caused by the spatial stretching effects inherent to MIP

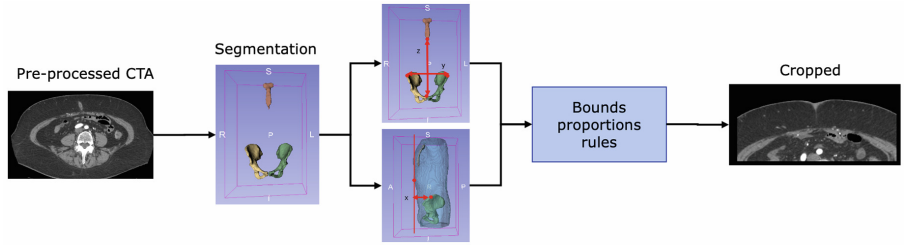


Fig. 3. A schematic overview of the anatomy-aware ROI cropping methodology

processes non-overlapping voxel patches and supplies multi-scale tokens to a convolutional decoder via skip connections. All dropout is disabled.

During training, four 96^3 patches extracted from cropped CTA ROI are sampled per iteration. To address the class imbalance (perforator vs background), three patches are centered on vessel voxels and one on background. Optimization is driven by a Dice + cross-entropy (CE) composite loss that balances region overlap with voxel-wise precision:

$$\mathcal{L} = 0.8\,\mathcal{L}_{\text{Dice}} + 0.2\,\mathcal{L}_{\text{CE}} \tag{1}$$

Augmentations (30% probability each) include sagittal flips, $\pm 5.7°$ rotations, $\pm 5\%$ zoom, Gaussian blur, gamma and intensity shifts, and Gaussian noise, thereby mimicking scanner and patient variability without distorting anatomy.

Models are trained with AdamW ($\eta_0 = 10^{-4}$, weight decay 10^{-5}). A five epoch linear warm-up is followed by cosine annealing to $\eta_{\min}$ over 200 epochs, with early stopping after 20 stagnant epochs.

Method 1 Proposed: Continuity-Aware Loss. To further enforce vessel continuity, the baseline loss is modified to directly penalize fragmented and topologically incorrect predictions $\mathcal{L} = 0.5\mathcal{L}_{\text{Dice}} + 0.3\mathcal{L}_{\text{clDice}} + 0.2\mathcal{L}_{\text{CE}}$, where the centerline Dice loss (clDice) [21] term enforces centerline agreement and thus penalizes breaks.

Method 2 Proposed: Region-aware Multi-class Training. The segmentation of perforators located within subcutaneous fat poses the greatest challenge because vessels appear low-contrast, thin and fragmented. To bias the network toward this under-represented context, the task is re-framed as two-class segmentation (perforators located in fat versus perforators located in other tissue) and optimizes the same loss with class weights that up-weight perforator surrounded by fat voxels. Fat masks are obtained automatically by applying the pre-trained CT body-composition model [5] and then combined with the original annotation to yield a multi-class mask. During training, patches are sampled stratified by these region labels, forcing the model to learn discriminative features in fat.

Postprocessing. Probability maps were binarized at $t = 0.5$. Components <10 voxels were discarded, and only the two largest "anchor" trunks were retained. A 3D vessel skeleton was then extracted from each retained component (both the anchor trunks and the remaining vessels) and smoothed with a B-spline to generate continuous centerlines. For bridge insertion, only tip voxels of pruned (non-anchor) components were considered: each tip within 4 voxels of either an anchor trunk or another pruned component was reconnected via a straight-line segment, preventing spurious trunk-to-trunk shortcuts. A closing operation (kernel = 2) filled minor holes, and any remaining component <50 voxels were removed, yielding continuous, anatomically realistic vessel masks.

Validation and Statistical Analysis

Train-test Split. The following train-test split was applied to segmentation methodology: (1) the baseline model hyper-parameters were first fine-tuned against 5 separated validation volumes, (2) the 5-fold cross-validation was applied to the remaining 50 volumes, allowing to each volume serving as the test case exactly once, thereby allowing metric evaluation on 50 patients.

Evaluation Metrics. Segmentation quality was assessed using standard volume and surface similarity metrics. These included: the Dice coefficient, voxel-wise volumetric Precision and Recall, and the 95th-percentile Hausdorff distance (HD95). Additionally, the centerline Dice - a topology-aware metric specifically designed for evaluating thin, connected structures like vessels and defined as:

$$\text{clDice} = \frac{2 \cdot T_p \cdot P_p}{T_p + P_p}, \tag{2}$$

where $T_p = \frac{|C(G) \cap P|}{|C(G)|}$ and $P_p = \frac{|C(P) \cap G|}{|C(P)|}$, and $C(P)$ and $C(G)$ represent the centerline (skeleton) extracted from the mask (or ground-truth) using morphological thinning, was employed.

Statistical Analysis. As validation metric distributions deviated from normality, a two-sided Wilcoxon signed-rank tests were applied to compare the baseline against each enhancement at a significance level of $\alpha = 0.05$, with Bonferroni correction to control the family-wise error.

3 Results

Manual Annotation with Depth-aware Back-Projection. The proposed depth-aware back-projection MIP annotation strategy substantially reduced manual annotation time by approximately 60–70%, from around 4 h down to about 90 min per patient, while significantly improving vessel continuity and decreasing the number of missed perforators. By using MIP images with spatial back-projection, the method allowed for the improved manual reading of perforator locations without impact of MIP-inherent spacing artifacts.

ROI Cropping Results. On ($n = 55$) CTA volumes, landmark detection and ROI cropping succeeded in 100% cases upon visual inspection, with a mean inference time of 28 ± 5 s per volume.

Perforator Segmentation Results. A comparison of the baseline Swin UNETR model against proposed enhancements (continuity-aware loss and region-aware multi-class training) was performed on 50 CTA volumes. Quantitative results are reported in Table 1, which lists median and interquartile range for Dice, clDice, Precision, Recall, and HD95. Figure 4 presents combined boxplots of all validation metrics. Figure 5 presents 3D renderings of predicted vessel masks, where the baseline model predicts accurately the main and some secondary perforator branches but exhibits fragmented distal branches. The continuity-aware loss restores some uninterrupted vessel trajectories, and the region-aware multi-class model more accurately delineates vessels within adipose tissue.

Table 1. Quantitative comparison of the baseline Swin UNETR vs proposed segmentation methods. All values median [Q_1, Q_3]

Model	Dice	Precision	Recall	HD95 (mm)	cl Dice
Swin UNETR	0.58 [0.54, 0.63]	0.65 [0.60, 0.70]	0.54 [0.47, 0.60]	20.0 [15.0, 23.8]	0.41 [0.36, 0.45]
Continuity-aware loss	0.60* [0.54, 0.65]	0.65* [0.58, 0.70]	0.58* [0.51, 0.44]	17.5 [11.1, 23.4]	0.41* [0.38, 0.46]
Region-aware multi-class	0.60* [0.56, 0.63]	0.63* [0.58, 0.70]	0.59* [0.52, 0.63]	19.15 [12.2, 24.7]	0.41 [0.35, 0.44]

* $p < 0.05$ vs. Swin UNETR.

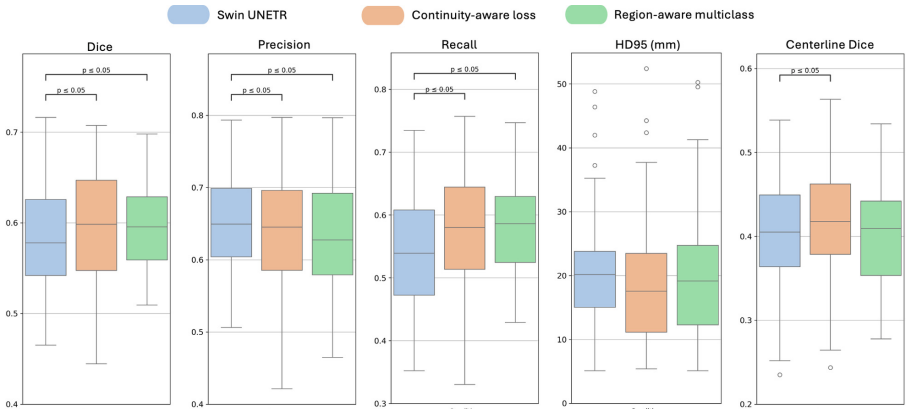


Fig. 4. Boxplots of segmentation performance across all 50 cases: Dice, Precision, Recall, HD95, and clDice for Swin UNETR (blue), continuity-aware loss (orange), and region-aware multi-class (green). Brackets indicate significant differences ($p \leq 0.05$) (Color figure online)

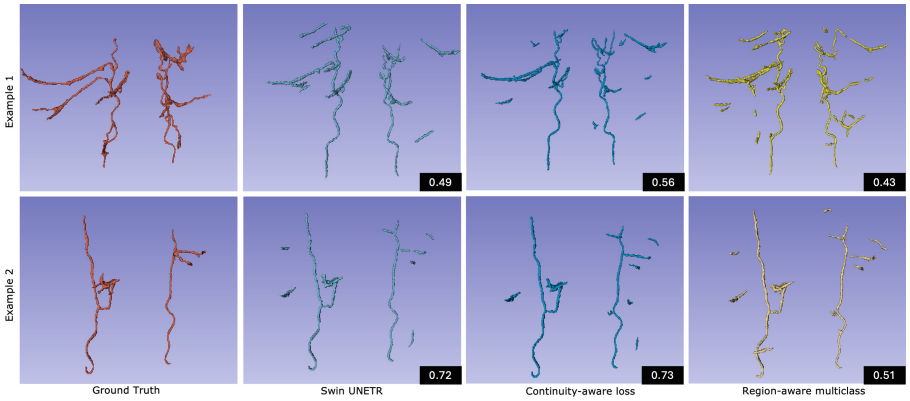


Fig. 5. A 3D rendering of a perforator segmentation results for a baseline and proposed segmentation methodologies on two cases (top: highly branched perforator; bottom: simpler anatomy). Dice scores are displayed in the bottom-right of each prediction

4 Discussion and Conclusions

The proposed depth-aware back-projection MIP annotation method substantially enhanced the annotation workflow by effectively addressing inherent spatial distortions associated with thin-slab MIP images and reduced the manual annotation times by approximately 6070%. The anatomy-aware ROI cropping demonstrated robust performance, successfully isolating the relevant anatomical region across all evaluated cases, thus decreasing the computational burden and enhancing segmentation precision by restricting analysis to the DIEP ter-

ritory. These two advances collectively set a solid foundation for an automated perforator segmentation method.

The baseline Swin UNETR model was able to segment the primary perforator trunks and several secondary branches, but produced fragmented segmentations in the smaller distal branches and demonstrated poor recall in low-contrast adipose regions. This was reflected by a moderate Dice similarity coefficient (0.58), low clDice (0.41), and relatively high HD95 (20 mm). These limitations underscore the challenge of capturing fine, submillimetre vessels and highlight the risk of false negatives due to low contrast in adipose tissue, as well as false positives arising from surrounding structures in the same region.

The two proposed methods for enhancing perforator segmentation addressed these limitations in complementary ways (see Table 1, Fig. 4). The integration of a customized continuity-aware loss significantly improved global overlap (Dice coefficient of 0.60, $p < 0.05$), reduced boundary errors (HD95 decreased to 17.5 mm), and enhanced centerline coherence (clDice increased to 0.41, $p < 0.05$). Additionally, the region-aware multi-class training method achieved comparable performance, with a Dice score of 0.60 ($p < 0.05$), significantly improved recall (0.59, $p < 0.05$), and only a slight reduction in precision (0.63 versus 0.65 for the baseline). It also maintained competitive boundary accuracy (HD95 of 19.2 mm) and topological consistency (clDice of 0.41). Qualitative evaluation based on the 3D renderings (Fig. 5) confirmed that both methods produced more continuous and anatomically realistic perforator vessel segmentations, although minor fragmentation persisted in very small distal branches.

In this work, we presented the first fully-automated pre-operative planning framework for DIEP flap surgery that integrates depth-aware MIP annotation, anatomy-guided ROI cropping, and enhanced perforator segmentation using continuity-aware and region-aware training. Validated on clinical CTA datasets, the proposed method substantially improved annotation efficiency, segmentation accuracy, and anatomical fidelity compared to baseline approaches, enabling automated perforator mapping in under 10 min per scan. These findings show the potential of this integrated pipeline to standardise perforator selection, decrease surgical planning time, and improve clinical outcomes. Future work will focus on prospective validation by radiologists, automated perforator ranking, and adaptation to other perforator-based surgical procedures.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Araujo, R.J., et al.: Computer aided detection of deep inferior epigastric perforators in computed tomography angiography scans. Comput. Med. Imaging Graph. **77**, 101648 (2019)
2. Boer, V.B., et al.: Concordance between preoperative computed tomography angiographic mapping and intraoperative perforator selection for deep inferior epigastric artery perforator flap breast reconstructions. Gland Surg. **6**(6), 620 (2017)

3. Bonafastia, S.J., Steenbeek, L.M., Ulrich, D.J., Hummelink, S.: Global perspectives on practices and preferences in autologous free flap breast reconstruction: from flap selection to postoperative care a descriptive quantitative study. JPRAS open **43**, 169–179 (2025)
4. Chae, M.P., Hunter-Smith, D.J., Rozen, W.M.: Imaging and printing in plastic and reconstructive surgery part 1: established techniques. Australas. J. Plast. Surg. **2**(1), 15–28 (2019)
5. Chen, Y., et al.: Automated muscle and fat segmentation in computed tomography for comprehensive body composition analysis. arXiv preprint arXiv:2502.09779 (2025)
6. Dima, A.F., et al.: 3D arterial segmentation via single 2D projections and depth supervision in contrast-enhanced CT images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 141–151. Springer (2023)
7. Fedorov, A., et al.: 3d slicer as an image computing platform for the quantitative imaging network. Magn. Reson. Imaging **30**(9), 1323–1341 (2012)
8. Garcia-Etienne, C.A., et al.: Mastectomy trends for early-stage breast cancer: a report from the eusoma multi-institutional european database. Eur. J. Cancer **48**(13), 1947–1956 (2012)
9. Hamdi, M., Kapila, A.K., Waked, K.: Current status of autologous breast reconstruction in europe: how to reduce donor site morbidity. Gland Surg. **12**(12), 1760 (2023)
10. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D.: Swin unetr: swin transformers for semantic segmentation of brain tumors in MRI images. In: International MICCAI Brainlesion Workshop, pp. 272–284. Springer (2021)
11. Hong, S.E., Kang, D.: Navigating the pandemic: shifts in breast reconstruction trends and surgical decision-making in the united states. J. Clin. Med. **13**(14), 4168 (2024)
12. International Agency for Research on Cancer: GLOBOCAN 2022: World Fact Sheet.https://gco.iarc.who.int/media/globocan/factsheets/populations/900-world-fact-sheet.pdf (2022), Accessed 03 May 2025
13. Issa, C.J., et al.: Comparing plastic surgeon operative time for DIEP flap breast reconstruction: 2-stage more efficient than 1-stage? Plast. Reconstr. Surg.-Global Open **9**(6), e3608 (2021)
14. Lange, C.J., et al.: Automating perforator flap mra and cta reporting. J. Digit. Imaging **30**, 350–357 (2017)
15. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440 (2015)
16. Nahabedian, M.Y.: The deep inferior epigastric perforator flap: where we started and where we are now. Gland Surg. **12**(5), 696 (2023)
17. Oliva, I.B., et al.: ACR appropriateness criteria® imaging of deep inferior epigastric arteries for surgical planning (breast reconstruction surgery). J. Am. Coll. Radiol. **14**(11), S456–S461 (2017)
18. Ronen, G., Blachar, A., Abelow, A., Gur, E., Barnea, Y.: Computed tomography angiography (CTA) assisted preoperative planning and volume calculation of deep inferior epigastric artery perforator (DIEP) flap for breast reconstruction. Eur. J. Plast. Surg. **40**(5), 441–446 (2017). https://doi.org/10.1007/s00238-017-1295-x
19. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi,

A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
20. Saxena, E.: Deep learning for personalized preoperative planning of microsurgical free tissue transfers. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, pp. 13140–13141 (2022)
21. Shit, S., et al.: clDice-a novel topology-preserving loss function for tubular structure segmentation. In: Proceedings of the IEEE/CVF Conference on Computer vision and Pattern Recognition, pp. 16560–16569 (2021)
22. Tetteh, G., et al.: Deepvesselnet: vessel segmentation, centerline prediction, and bifurcation detection in 3-d angiographic volumes. Front. Neurosci. **14**, 592352 (2020)
23. Wasserthal, J., et al.: Totalsegmentator: robust segmentation of 104 anatomic structures in CT images. Radiol. Artif. Intell. **5**(5), e230024 (2023)

MM + CD Fusion: Deep Learning-based 3D Multi-Modal Fusion for Early Pathological Complete Response Prediction in Breast Cancer

Tewele W. Tareke[1(✉)], Neree Payan[1,2], Alexandre Cochet[1,2], Alain Lalande[1,3], and Fabrice Meriaudeau[1]

[1] ICMUB Laboratory, UMR CNRS 6302, Université de Bourgogne Europe, 21000 Dijon, France
tewele-weletnsea.tareke@u-bourgogne.fr
[2] Nuclear Medicine Department, Centre Georges-François Leclerc, 21000 Dijon, France
[3] Department of Medical Imaging, University Hospital of Dijon, 21000 Dijon, France
https://icmub.ube.fr/
https://www.cgfl.fr
https://www.chu-dijon.fr/

Abstract. This study aims to evaluate the predictive potential of baseline ^{18}F-Fluorodeoxyglucose Positron Emission Tomography/Computed Tomography (^{18}F-FDG PET/CT), clinical data (CD), dynamic contrast-enhanced MRI (DCE-MRI), and T1-weighted MRI (T1w-MRI) parameters, including advanced texture features, for assessing pathological complete response (pCR) to neoadjuvant chemotherapy (NAC) in patients with advanced breast cancer (BC), using a cohort of 81 BC patients. Accurate identification of patients who achieve pCR to NAC is crucial for guiding personalized treatment strategies and assessing prognosis in breast cancer. However, most existing pCR prediction methods rely on single-modality data, which limits their ability to capture tumor dynamics and fully represent tumor heterogeneity at both macro and micro levels. Moreover, these approaches often fail to effectively integrate the complementary information available across imaging modalities. In this study, we propose **MM (multi-modal) + CD Fusion**, a framework that pioneers the integration of MM imaging and CD for enhanced treatment response prediction. Our approach consists of two core components: (1) MM imaging, which includes ^{18}F-FDG PET/CT and multiparametric MRI (DCE and T1w), offering comprehensive metabolic and anatomical representations of breast cancer lesions and their NAC-induced changes; and (2) CD, which incorporates clinical variables characterizing tumor biology. We introduce label propagation techniques to enable automatic lesion segmentation across all MM images using a 3D U-Net deep learning model, followed by pCR prediction through machine learning algorithms. The proposed **MM + CD fusion** approach significantly outperforms unimodal methods in predicting pCR (AUC: 0.87 vs. 0.63–0.77, $p < 0.012$) using 141 optimally selected radiomic features. This improve-

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
T. Zhang et al. (Eds.): Deep-Breath 2025, LNCS 16142, pp. 134–144, 2026.
https://doi.org/10.1007/978-3-032-05559-0_14

ment supports the potential of MM + CD fusion for more personalized therapy planning in breast cancer patients.

Keywords: Breast Cancer · Deep learning · Segmentation · Radiomic features · Machine learning · pCR prediction · NAC

1 Introduction

Breast cancer remains a leading cause of cancer-related mortality among women. Neoadjuvant chemotherapy (NAC) is widely used for managing advanced breast cancer, but its effectiveness can only be confirmed through post-surgical pathological examination [1]. Patients achieving a pathological complete response (pCR) to NAC typically have improved survival and lower recurrence rates [2]. However, less than 30% of patients reach pCR [3], and early identification of non-responders is critical to adapt treatment strategies more effectively. Breast cancer subtypes exhibit varied responses to therapy, necessitating personalized approaches [4]. Imaging plays a vital role in monitoring NAC response, with modalities such as fluorodeoxyglucose positron emission tomography/computed tomography (FDG-PET/CT), dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI), T1-weighted magnetic resonance imaging (T1w-MRI) and mammography. PET/CT assesses metabolic changes [5], while MRI evaluates structural and vascular alterations-both of which are key indicators of therapy response [6]. As a result, ^{18}F-FDG PET/CT and MRI are increasingly used to predict treatment outcomes prior to surgery [7]. Multi-modal (MM) imaging techniques are commonly used to obtain more comprehensive information than a single modality can provide, as studied by Ali et al. [8]. Recent research highlights the potential of combining multiparametric imaging with artificial intelligence to improve pCR prediction accuracy [9]. Zhang et al. [10] provided a comprehensive review of recent advances, challenges, and the future impact of radiomics and AI in breast cancer diagnosis and research. Deep learning models, particularly convolutional neural networks (CNNs), have shown promise in cancer detection, classification, and outcome prediction across various imaging modalities [11]. Although CNNs have been applied in several cancer types [12], their use in predicting pCR from PET/CT in breast cancer remains limited [13]. While machine learning (ML) techniques like random forests have been effective in extracting radiomic features that correlate with treatment response [14], their performance is often limited by the isolated analysis of single imaging modalities (e.g., PET or MRI alone). By integrating anatomical, metabolic, and clinical information through fusion, predictive power of pCR can be further enhanced [15].

To address existing challenges, we propose integrating MM imaging-comprising ^{18}F-FDG PET/CT, DCE-MRI, and T1w-MRI-with clinical data (CD) to improve pCR prediction through the combined interpretation of imaging and clinical information. The key contributions of this work are:

1. We introduce a label propagation approach to transfer labels across unlabeled imaging modalities, enabling robust representation extraction from automatically segmented lesions.

2. We present **MM + CD Fusion**, a novel framework that integrates radiomic and metabolic features from multiple imaging modalities with clinical data (CD), enhancing the exploitation of complementary information for predicting pCR.
3. To the best of our knowledge, this is the first study to simultaneously integrate MM imaging and clinical data for treatment response prediction. Our approach achieves improved predictive performance while maintaining strong generalization across the dataset.

2 Materials and Methods

2.1 Data Collection and Patient Stratification

We analyzed data from 254 patients enrolled in a study at the Anti-Cancer Center Georges-François Leclerc (CGFL), Dijon, France, with multiple imaging modalities available [16]. A retrospective subset of 92 patients underwent PET/CT and MRI scans (2011–2017) at the University Hospital Center in Dijon Bourgogne. Baseline ^{18}F-FDG PET/CT (Philips Medical Systems, Eindhoven, The Netherlands) was acquired before NAC, followed by post-treatment scans. Patients received 3 MBq/kg ^{18}F-FDG, with an 8-minute prone breast-centered PET scan and low-dose CT for attenuation correction. MRI (1.5 T magnet, Siemens Healthineers, Erlangen, Germany) included DCE subtraction imaging with T1w turbo spin-echo sequences. Molecular subtypes (luminal, triple-negative, HER2) were determined via biopsy, assessing hormone receptors (ER/PR) and HER2 status [17]. Of 92 patients, 81 completed follow-up, with 32 classified as responders and 49 as non-responders.

2.2 Imaging Alignment and Automated Segmentation

We present a fully automated MM imaging and clinical data (MM + CD) fusion pipeline for early prediction of treatment response in breast cancer. An overview of the pipeline is illustrated in Fig. 1. The pipeline integrates PET/CT and DCE-MRI/T1w scans with corresponding CD. PET and DCE-MRI images were annotated by experts, whereas CT and T1w-MRI scans remained unlabeled. To address this, we proposed a label propagation strategy to transfer annotations to CT and T1w-MRI, enabling automated segmentation across all imaging modalities. PET and CT, as well as DCE-MRI and T1w-MRI, were co-registered during acquisition for all patients. However, differences in image dimensions and voxel spacing across modalities can introduce inconsistencies. We propagated (PG) the ground truth labels from PET to CT and from DCE-MRI to T1w-MRI via resampling to harmonize voxel spacings(DCE-MRI: $0.86 \times 0.86 \times 1.0\,\text{mm}^3$; T1w-MRI: $0.75 \times 0.75 \times 3.0\,\text{mm}^3$; PET/CT: $4.0 \times 4.0 \times 4.0\,\text{mm}^3/1.17 \times 1.17 \times 2.50\,\text{mm}^3$). The resampling process to ensure spatial consistency across modalities is explained as follows:

Modality-Specific Resampling (Resam)

PET → CT:

$$\text{Resam_factor}_d^{\text{PET}\to\text{CT}} = \frac{\text{PET_dim}_d \cdot 4.0}{\text{CT_dim}_d \cdot \text{CT_spacing}_d}$$

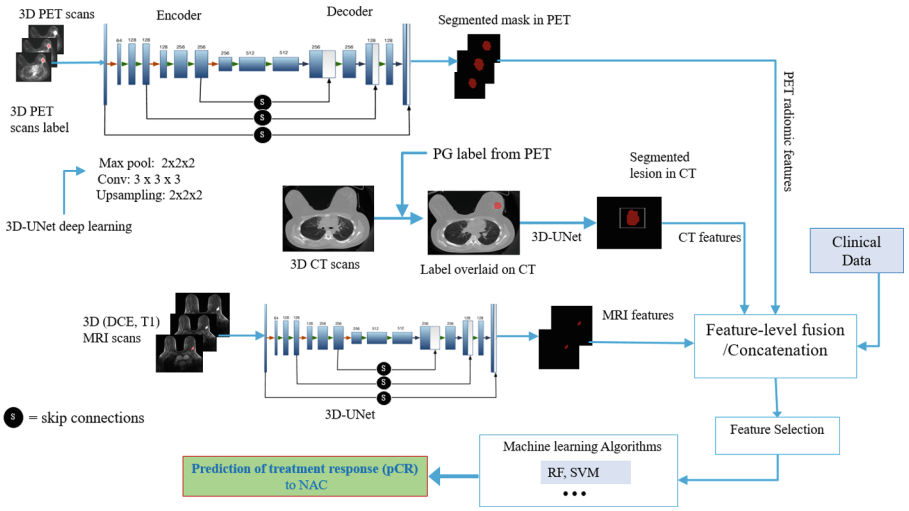


Fig. 1. Pipeline of the proposed study integrating MM image segmentation, feature extraction, and CD fusion for pCR prediction to NAC

where PET_dim$_d$ is the voxel count in PET, CT_dim$_d$ and CT_spacing$_d$ are the voxel count and spacing in CT, and 4.0 mm is the default PET voxel spacing.

DCE-MRI → T1w-MRI:

$$\text{Resam_factor}_d^{\text{DCE}\to\text{T1w}} = \frac{\text{DCE_dim}_d \cdot \text{DCE_spacing}_d}{\text{T1w_dim}_d \cdot \text{T1w_spacing}_d}$$

Validation Metrics for Label Propagation: In addition to visual inspection by overlaying the propagated labels on CT and DCE-MRI images, we performed the following quantitative validations.

1. Volume Preservation Check (Vol. Err. (%)) Given a binary mask in source space L_{src} and its propagated mask L'_{tgt}, physical volume should remain consistent:

PET → CT:

$$\text{Vol}_{\text{PET}} = \sum \mathbf{1}(L_{\text{PET}} = 1) \cdot 4.0^3$$

$$\text{Vol}_{\text{CT}} = \sum \mathbf{1}(L'_{\text{CT}} = 1) \cdot s_x^{\text{CT}} s_y^{\text{CT}} s_z^{\text{CT}}$$

$$\Delta V = \frac{|\text{Vol}_{\text{PET}} - \text{Vol}_{\text{CT}}|}{\text{Vol}_{\text{PET}}}$$

DCE-MRI → T1w-MRI:

$$\text{Vol}_{\text{DCE}} = \sum \mathbf{1}(L_{\text{DCE}} = 1) \cdot s_x^{\text{DCE}} s_y^{\text{DCE}} s_z^{\text{DCE}}$$

$$\text{Vol}_{\text{T1}} = \sum \mathbf{1}(L'_{\text{T1}} = 1) \cdot s_x^{\text{T1}} s_y^{\text{T1}} s_z^{\text{T1}}$$

$$\Delta V = \frac{|\text{Vol}_{\text{DCE}} - \text{Vol}_{\text{T1}}|}{\text{Vol}_{\text{DCE}}}$$

If $\Delta V < 5\%$, volume preservation is considered acceptable.

2. Dice Similarity Round-Trip Projection (DRTP)

$$\textbf{PET} \rightarrow \textbf{CT} \rightarrow \textbf{PET:} \quad \text{Dice} = \frac{2|L_{\text{PET}} \cap L_{\text{PET}}^{\text{roundtrip}}|}{|L_{\text{PET}}| + |L_{\text{PET}}^{\text{roundtrip}}|}$$

$$\textbf{DCE} \rightarrow \textbf{T1w} \rightarrow \textbf{DCE:} \quad \text{Dice} = \frac{2|L_{\text{DCE}} \cap L_{\text{DCE}}^{\text{roundtrip}}|}{|L_{\text{DCE}}| + |L_{\text{DCE}}^{\text{roundtrip}}|}$$

DRTP ≥ 0.95 indicates reliable round-trip consistency; lower values may suggest interpolation or misalignment issues.

We trained separate 3D U-Net models on CT and T1w-MRI volumes using pseudo-labels propagated from PET and DCE-MRI, respectively, within a supervised training setup. This enables cross-modality knowledge transfer, where high-resolution anatomical structures are learned from the corresponding functional source modalities. As a result, the framework achieved generalizable segmentation and pCR prediction performance across CT and T1w-MRI in single-modality settings, even when PET or DCE-MRI are unavailable during inference. This leads to a fully automated framework that is more adaptable to real-world clinical applications, where access to PET and DCE-MRI imaging may be limited. In summary, we developed both single- and MM fused computer-aided prediction models across all imaging modalities by initially leveraging label propagation techniques. These models can later be applied to new CT and T1w-MRI cases without the need for further label propagation.

2.3 CD and Multi-Modal Feature Extraction

From the aligned ROIs, we extracted a comprehensive set of 272 radiomic features, including first-order statistics (entropy, kurtosis, SUVmax, SUVmean, metabolic tumor volume (MTV), and total lesion glycolysis), texture features computed using Gray-Level Co-occurrence Matrices (GLCM) across 13 spatial directions (3D offsets at inter-voxel distance = 1) that preserved directional information for four key metrics (contrast, homogeneity, correlation, energy) totaling 52 features per modality, along with morphological characteristics (sphericity, compactness, margin sharpness, elongation). These were complemented with 11 clinical parameters: T stage, N stage, Scarff Bloom Richardson (SBR) grade, SBR mitotic count, age, estrogen receptor (ER) status, progesterone receptor (PR) status, HER2 status, histological type, menopausal status, and tumor subtype classification (luminal, TN, HER2) [17]. The complete feature set (66 from PET ($\boldsymbol{f}_{\text{PET}} \in \mathbb{R}^{66}$), 130 from MRI (DCE, T1w) ($\boldsymbol{f}_{\text{MRI}} \in \mathbb{R}^{130}$), 11 from clinical data ($\boldsymbol{f}_{\text{CD}} \in \mathbb{R}^{11}$), and 65 from CT ($\boldsymbol{f}_{\text{CT}} \in \mathbb{R}^{65}$)) provided a multi-dimensional representation of tumor biology, integrating functional, textural, morphological, and clinical biomarkers for comprehensive response prediction. To ensure comparable scales across modalities, we apply *z-score* normalization:

$$\tilde{\boldsymbol{f}}_i = \frac{\boldsymbol{f}_i - \mu_i}{\sigma_i}, \quad \text{for } i \in \{\text{PET}, \text{CD}, \text{MRI}, \text{CT}\}, \tag{1}$$

where μ_i and σ_i are the mean and standard deviation of features for modality i.

The normalized features are combined into a single high-dimensional vector:

$$\boldsymbol{F}_{\text{fused}} = \left[\tilde{\boldsymbol{f}}_{\text{PET}} \| \tilde{\boldsymbol{f}}_{\text{CD}} \| \tilde{\boldsymbol{f}}_{\text{MRI}} \| \tilde{\boldsymbol{f}}_{\text{CT}}\right] \in \mathbb{R}^{d_1+d_2+d_3+d_4}, \tag{2}$$

where $\|$ denotes vector concatenation.

2.4 Implementation and Evaluation

Separate 3D U-Nets were trained to segment lesions in ^{18}F-FDG PET and DCE-MRI using 5-fold cross-validation on 81 subjects, ensuring each subject appeared in the validation set once. Each U-Net had 5 levels and was trained using the Adam optimizer (weight decay = 1×10^{-4}, initial learning rate = 1×10^{-3}), with Leaky ReLU activations and batch normalization in all convolutional blocks.The method was developed using PyTorch and an NVIDIA RTX 3090 GPU with 24 GB of VRAM. A fixed random seed (42) was used. Segmentation was evaluated using Dice score, Intersection over Union (IoU), and Hausdorff distance (HD). Radiomic features from MM imaging were extracted, standardized, and combined, which selected 139 out of the 272 fused features. We employed recursive feature elimination to select the most discriminative features, enhancing the model's performance and interpretability. For prediction, an SVM (RBF kernel) and a Random Forest (500 trees, max depth = 15) were optimized via grid search. Performance was reported using stratified 5-fold CV with AUC, F1-score, and balanced accuracy.

3 Result and Discussion

3.1 Label Propagation and Segmentation

The label propagation achieved excellent DRTP scores (>0.94) across all cases, with volume errors consistently below 3.5% (see Table 1). Automatic segmentation achieved Dice scores of 0.79±0.04 for PET, 0.78 ± 0.03 for DCE-MRI, 0.68 ± 0.01 for CT, and 0.71 ± 0.03 for T1w-MRI. The binary segmentation results are visually illustrated in the segmented masks (Fig. 2). Among all modalities, PET and DCE-MRI demonstrated the highest accuracy (Dice/IoU) and the lowest HD. The superior segmentation performance of PET and DCE-MRI may be attributed to their high tumor-to-background contrast and functional imaging capabilities, which provide more distinct tumor boundaries compared to anatomical modalities like CT or T1w-MRI.

3.2 pCR Prediction

In the combined analysis of all breast cancer subtypes, the MM + CD approach achieved the highest predictive performance (Bacc: 0.82 ± 0.02, AUC: 0.87 ± 0.07). Individual modalities yielded lower results, with MRI (AUC: 0.76 ± 0.02) and PET (AUC: 0.74 ± 0.02) performing comparably, while CD (AUC: 0.63 ±

Table 1. Label propagation validation metrics (5 sample cases + average of 81 cases)

Case #	PET → CT		DCE-MRI → T1w-MRI	
	Vol. Err. (%)	DRTP	Vol. Err. (%)	DRTP
1	2.1	0.97	2.5	0.96
2	3.0	0.96	2.8	0.95
3	1.8	0.98	2.2	0.97
4	2.5	0.95	2.9	0.94
5	1.7	0.97	2.0	0.96
Avg. (81 cases)	1.8 ± 0.4	0.97 ± 0.01	2.1 ± 0.6	0.96 ± 0.03

Note: Volume Error (Vol. Err.) measures physical consistency after propagation; DRTP (Dice Round-Trip Projection) tests label fidelity. Values reported as mean ± standard deviation for the remaining 81 cases.

Table 2. Performance comparison of MM + CD fusion and unimodal methods for predicting pCR (mean ± std, 5-fold cross-validation) for luminal + HER2 + TN BC subtypes

Metric	PET	CT	T1	DCE	MRI	CD	MM	CD+MM
Bacc	0.74 ± 0.05	0.66 ± 0.02	0.71 ± 0.07	0.73 ± 0.03	0.76 ± 0.03	0.61 ± 0.04	0.83 ± 0.05	$\mathbf{0.85 \pm 0.02}$
F1	0.70 ± 0.07	0.56 ± 0.06	0.68 ± 0.01	0.71 ± 0.07	0.72 ± 0.01	0.60 ± 0.07	0.77 ± 0.02	$\mathbf{0.79 \pm 0.05}$
AUC	0.74 ± 0.02	0.66 ± 0.06	0.72 ± 0.04	0.74 ± 0.01	0.76 ± 0.07	0.63 ± 0.01	0.84 ± 0.07	$\mathbf{0.87 \pm 0.07}$
Recall	0.67 ± 0.06	0.56 ± 0.02	0.65 ± 0.07	0.68 ± 0.03	0.69 ± 0.03	0.59 ± 0.04	0.74 ± 0.06	$\mathbf{0.82 \pm 0.06}$
Prec.	0.60 ± 0.04	0.57 ± 0.07	0.63 ± 0.04	0.62 ± 0.08	0.69 ± 0.06	0.58 ± 0.08	0.72 ± 0.02	$\mathbf{0.80 \pm 0.02}$

Note: Best in bold. MRI = T1+DCE; MM = Multi-modal (PET + CT + MRI); CD = Clinical Data.

0.01) and CT (AUC: 0.66 ± 0.06) showed the weakest performance (Table2). The fusion of PET, CT, MRI (DCE, T1w), and CD features significantly improved prediction accuracy, as evidenced by the combined ROC curve (Fig. 2). This underscores the complementary value of integrating anatomical and functional imaging with clinical data for enhanced lesion characterization and outcome prediction.

This study aimed to predict pCR before the first cycle of NAC in breast cancer patients by integrating metabolic and radiomic features from PET/CT and multiparametric MRI (DCE, T1-weighted), along with clinical-demographic data. Our proposed approach employed a MM imaging framework that combines radiomic, metabolic, and CD features across diverse breast cancer molecular subtypes to construct a robust baseline-level prediction model. The methodology was designed to address sample size limitations while effectively leveraging the complementary strengths of MM imaging and CD information to distinguish pCR from non-pCR cases. Deep learning-based features were extracted from automatically segmented primary tumors on baseline PET/CT and MRI scans,

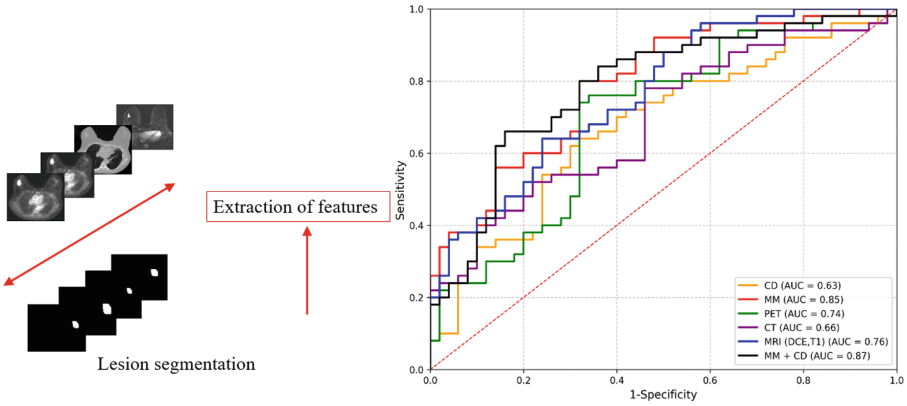


Fig. 2. Automated MM Lesion segmentation, Feature extraction with pCR prediction ROC analysis (unimodal vs. MM Imaging)

then fused with CD features. This fusion model significantly outperformed both unimodal and CD-only models ($p < 0.014$). To the best of our knowledge, this is the first study demonstrating that automatic integration of PET/CT, MRI, and CD features from segmented lesions enhances early pCR prediction, representing a notable step forward in personalized treatment planning. Feature selection and importance ranking were performed using mutual information and SHAP values, respectively, enabling identification of consistently predictive features while maintaining model interpretability and efficiency. Automatic segmentation achieved high accuracy (Dice: PET 0.79 ± 0.05, DCE-MRI 0.78 ± 0.02) with superior boundary delineation (lower HD), while multi-modal integration (PET/CT/MRI) yielded the best pCR prediction (Bacc: 0.85 ± 0.05, AUC: 0.87 ± 0.07), outperforming single modalities by 7–21%. While prior studies have investigated machine learning approaches for pCR prediction [18], automated systems for lesion segmentation across full MM imaging remain limited [19]. Ypsilantis et al. [20] used CNNs with FDG-PET to evaluate NAC response, achieving an accuracy of 0.73 ± 0.05 versus 0.57 ± 0.07 with RF, but lacked anatomical integration, resulting in imprecise tumor boundaries due to PET's low spatial resolution. To overcome this, we developed a method combining PET, CT, MRI, and CD features, capturing complementary information including tumor margins, and achieved 0.85 ± 0.02 Bacc with MM imaging + CD. Cha et al. [21] used CNNs on pre-treatment CT in rectal cancer, reporting an AUC of 0.73. While our CT-only model yielded a lower AUC (0.66 ± 0.06), integrating MM imaging + CD significantly improved performance (AUC = 0.87 ± 0.07). Payan et al. [17] reported a Bacc of 0.66±0.09 for the HER2 + TN subtypes using baseline PET and clinical data. In comparison, our method achieved a higher Bacc of 0.77±0.08 on the HER2 + TN subtypes, despite being trained on a smaller dataset (62 vs. 128 cases). The integration of MM imaging with CD further improved performance, reaching a Bacc of 0.85±0.02. Similarly, Lim et al.

[22] reported PET-based AUCs of 0.73 (PET alone) and 0.77 (PET + clinical), while our MM + CD approach yielded superior performance. Future improvements could involve expanding the training dataset with additional PET/CT and MRI images linked to clinical and genomic data. We also plan to develop a multi-modal imaging algorithm that incorporates histopathological data and T2-weighted MRI, which were not explored in this study. Integrating whole-slide histopathology with FDG-PET, CT, and MRI (DCE, T1w, T2-weighted MRI) data is expected to enhance advanced NAC response prediction. Such models could improve preoperative patient stratification for clinical trials. While our results are promising, further validation is required. We are currently addressing this by assembling an independent cohort from an external institution to verify the generalizability of our fusion model.

4 Conclusion

This study presents a novel MM + CD approach for predicting pCR before the first cycle of NAC in breast cancer patients, integrating radiomic and metabolic features from PET/CT and multiparametric MRI (DCE, T1w), along with clinical-demographic data. By combining deep learning-based features extracted from automatically segmented tumors with structured CD information, the proposed model achieved significantly higher predictive performance than unimodal or CD-only baselines. Our method addresses limitations in prior work, which often lacked anatomical integration or relied on single modalities with lower spatial resolution. Through comparative analysis with state-of-the-art studies, we demonstrate consistent improvements across HER2, TN, and luminal subtypes(even with a smaller dataset), highlighting the strength of our multi-modal design.

Acknowledgments. This study was funded by Marie Skodowska-curie doctoral network actions (horizon-msca-2021-dn-01-01) with a grant number of 101073222.

Disclosure of Interests. The authors declare no conflict of interest.

References

1. Park, C.K., Jung, W.-H., Koo, J.S.: Pathologic evaluation of breast cancer after neoadjuvant therapy. J. Pathol. Transl. Med. **50**(3), 173–180 (2016)
2. Rastogi, P., et al.: Preoperative chemotherapy: updates of national surgical adjuvant breast and bowel project protocols b-18 and b-27. J. Clin. Oncol. **26**(5), 778–785 (2008)
3. Heller, W., et al.: Neoadjuvant 5-fluorouracil, epirubicin and cyclophosphamide chemotherapy followed by docetaxel in refractory patients with locally advanced breast cancer. Oncol. Rep. **17**(1), 253–259 (2007)
4. Veronesi, U., Goldhirsch, A., Veronesi, P., Gentilini, O.D., Leonardi, M.C.: Breast cancer: Innovations in research and management. Springer (2017)

5. Sheikhbahaei, S., et al.: Fdg-pet/ct and mri for evaluation of pathologic response to neoadjuvant chemotherapy in patients with breast cancer: a meta-analysis of diagnostic accuracy studies. Oncologist **21**(8), 931–939 (2016)
6. Pengel, K.E., et al.: Combined use of 18 f-fdg pet/ct and mri for response monitoring of breast cancer during neoadjuvant chemotherapy. Eur. J. Nuclear Med. Molecul. Imaging **41**, 1515–1524 (2014)
7. Joye, I., Deroose, C.M., Vandecaveye, V., Haustermans, K.: The role of diffusion-weighted mri and 18f-fdg pet/ct in the prediction of pathologic complete response after radiochemotherapy for rectal cancer: a systematic review. Radiother. Oncol. **113**(2), 158–165 (2014)
8. Ali, Y., Gregori, J., Tareke, T.W., Lalande, A., Meriaudeau, F.: Leveraging mri radiomics and machine learning for accurate differentiation of triple-negative breast cancer subtype. In: IEEE 38th International Symposium on Computer-Based Medical Systems (CBMS). IEEE Computer Society, vol. 2025, pp. 922–928 (2025)
9. Pinker, K., et al.: Improved differentiation of benign and malignant breast tumors with multiparametric 18fluorodeoxyglucose positron emission tomography magnetic resonance imaging: a feasibility study. Clin. Cancer Res. **20**(13), 3540–3549 (2014)
10. Zhang, T., et al.: Radiomics and artificial intelligence in breast imaging: a survey. Artif. Intell. Rev. **56**(Suppl 1), 857–892 (2023)
11. Choi, J.H., et al.: Early prediction of neoadjuvant chemotherapy response for advanced breast cancer using pet/mri image deep learning. Sci. Rep. **10**(1), 21149 (2020)
12. Zhang, H., Han, L., Chen, K., Peng, Y., Lin, J.: Diagnostic efficiency of the breast ultrasound computer-aided prediction model based on convolutional neural network in breast cancer. J. Digit. Imaging **33**, 1218–1223 (2020)
13. Yamashita, R., Nishio, M., Do, R.K.G., Togashi, K.: Convolutional neural networks: an overview and application in radiology. Insights Imaging **9**(4), 611–629 (2018). https://doi.org/10.1007/s13244-018-0639-9
14. Shen, W.-C., et al.: Predicting pathological complete response in rectal cancer after chemoradiotherapy with a random forest using 18f-fluorodeoxyglucose positron emission tomography and computed tomography radiomics. Ann. Transl. Med. **8**(5), 207 (2020)
15. Li, P., et al.: 18 f-fdg pet/ct radiomic predictors of pathologic complete response (pcr) to neoadjuvant chemotherapy in breast cancer patients. Eur. J. Nucl. Med. Mol. Imaging **47**, 1116–1126 (2020)
16. Payan, N., et al.: Biological correlates of tumor perfusion and its heterogeneity in newly diagnosed breast cancer using dynamic first-pass 18 f-fdg pet/ct. Eur. J. Nucl. Med. Mol. Imaging **47**, 1103–1115 (2020)
17. Payan, N., et al.: Respective contribution of baseline clinical data, tumour metabolism and tumour blood-flow in predicting pcr after neoadjuvant chemotherapy in her2 and triple negative breast cancer. EJNMMI Res. **14**(1), 60 (2024)
18. Gullo, R.L., Eskreis-Winkler, S., Morris, E.A., Pinker, K.: Machine learning with multiparametric magnetic resonance imaging of the breast for early prediction of response to neoadjuvant chemotherapy. Breast **49**, 115–122 (2020)
19. Ha, R., et al.: Prior to initiation of chemotherapy, can we predict breast tumor response? deep learning convolutional neural networks approach using a breast mri tumor dataset. J. Digit. Imaging **32**, 693–701 (2019)
20. Ypsilantis, P.-P., et al.: Predicting response to neoadjuvant chemotherapy with pet imaging using convolutional neural networks. PLoS ONE **10**(9), e0137036 (2015)

21. Cha, K.H., et al.: Bladder cancer treatment response assessment in ct using radiomics with deep-learning. Sci. Rep. **7**(1), 8738 (2017)
22. Lim, C.H., et al.: Development and external validation of 18f-fdg pet-based radiomic model for predicting pathologic complete response after neoadjuvant chemotherapy in breast cancer. Cancers **15**(15), 3842 (2023)

Exploring Synergies Between Convolutional Neural Networks and Transformers for Breast Cancer Segmentation

Carlos Santiago and Jacinto C. Nascimento(✉)

Institute for Systems and Robotics, Instituto Superior Técnico, Lisbon, Portugal
jacinto.nascimento@tecnico.ulisboa.pt

Abstract. Breast cancer requires precise and early detection for effective treatment. Lesion segmentation is a pivotal stage to screen for the patient's condition and often integrated in CAD systems designed to assist clinicians. However, such systems still struggle with high false positive rates. Conventional segmentation methods mainly rely on convolutional neural networks (CNNs), proficient in local patterns but lacking in distant correlations. Recent advancements in Vision Transformers (ViTs), have shown improved efficacy and accuracy, as they integrate attention mechanisms, emphasizing finer details and global context. This work proposes a new mammography image segmentation model that combines the benefits of CNNs and Transformers, through an encoder module with two branches. Their information is then fused with an attention mechanism to help identify relevant spatial and channel features. The proposed model achieves state-of-the-art results in both the CBIS-DDSM and INbreast benchmark datasets, demonstrating the advantage of merging these techniques into a single model.

Keywords: Breast Cancer · Segmentation · Transformers · Encoder Fusion

1 Introduction

Breast cancer has recently become the most common type of cancer worldwide and is the leading cause of cancer-related death among women [2]. When detected at an early stage, treatment is highly effective and can significantly improve survival rates by identifying the disease before it metastasizes. However, the detection process is often time-consuming and labor-intensive, requiring prolonged observation by clinicians. Automatic segmentation methods, commonly integrated into computer-aided diagnosis (CAD) systems, have emerged as valuable tools for identifying breast lesions in medical images. These help ensure that clinicians do not overlook potentially important findings, thus improving their accuracy while reducing the time for image interpretation. Nonetheless,

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
T. Zhang et al. (Eds.): Deep-Breath 2025, LNCS 16142, pp. 145–154, 2026.
https://doi.org/10.1007/978-3-032-05559-0_15

their widespread adoption in routine screening remains controversial due to the high rate of false positives and the risk of misinterpretation, which means these models are not always cost-effective [3].

Convolutional neural networks (CNNs) are a powerful tool for tumor detection in digital pathology reaching impressive accuracies of over 90% for tumor detection depending on the type of cancer [8,19]. The architecture is purely convolutional, not capturing the correlation of regions in the image that are spatially distant. More recently, Transformers (*e.g.*, (ViTs) [16]) became a de-facto choice for modelling long-range dependencies. Compared to CNN methods, transformer models have larger receptive fields and excel at learning global information. In [16] it is shown that the entirely convolution-free in ViTs can outperform CNNs such as ResNet [18]. Also, it has been shown the versatility of transformers for natural language processing, concretely, BERT [15], RoBERTa [20], language models [22] or XLNet [27]. The findings above raise the question whether ViTs are the next milestone model in image recognition and whether those will replace CNNs in the future - the so called "Roaring 20 s" [21] that began with ViTs.

In medical image segmentation, several studies have proposed hybrid architectures that combine Transformers and CNNs to capture both local and global representations, for instance, TFCN [14], UNETR [17], SwinUNETR [1], TransUNet [5], MedT [25], CoTr [26]. Although achieving promising performances, these methods are still questionable since they incur several limitations. First, they have limited capability to learn both local and global representations due to ineffective integration of CNN and Transformer. For instance, CoTr [26] did not apply convolutions to their deformable Transformer for an effective learning strategy of local information. TransUNet [5] simply appended transformer to a convolutional block in the encoder. UNETR [17] did not account for a leverage convolution in the encoder. Nevertheless, and to summarize, the dichotomy of CNN vs ViTs remains an active topic of discussion. This still motivates research in the architectural distinctions between them. Namely, trying to identify the confounding variables when comparing the network performances, bridging the gap between the pre-ViT and post-ViT eras for CNNs, as well as testing the limits of what pure CNNs can accomplish [21].

This work proposes a new architecture that combines a CNN and a Transformer encoder in an innovative way. In particular, we borrow inspiration from the ONet [11] that divides itself into two parts. In the encoder part, a CNN is combined with the Swin Transformer, allowing to capture both global and local contextual features. In the decoder part, the results of the Swin Transformer and the CNN blocks are fused to obtain the final segmentation results. However, instead of fusing information directly, as originally proposed in ONet [11], we take inspiration from the TFCN model [14], strengthening the CNN branch with a Convolutional Linear Attention Block (CLAB) module before fusing its features with the transformer branch (see Fig. 1). This new architecture combines the benefits of CNNs, attention mechanisms, and Transformers, leading to

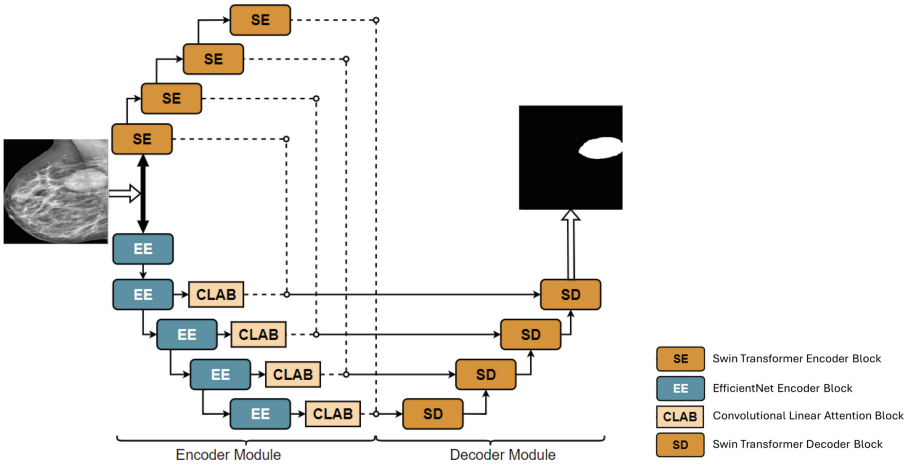


Fig. 1. The new proposed model architecture, which consists of an encoder with a CNN-based branch (top) and a Transformer-based branch (bottom), and CLAB attention blocks that fuse both branches before the decoder generates the segmentation

state-of-the-art results for breast cancer segmentation in mammography images in two popular benchmark datasets, CBIS-DDSM [9] and INbreast [4].

2 Proposed Method

The novel architecture (see Fig. 1) has two main branches: 1) one branch that uses Swin Transformer blocks, and 2) the second branch that employs Efficient Net blocks. The novelty presented herein consists in the introduction of a new feature fusion strategy that uses CLAB attention blocks in the EfficientNet branch before feeding the joint information to the decoder block. These three modules are described in the following subsections.

2.1 Swin Transformer Encoder

Different from the transformer, Swin transformer is built based on shifted windows rather than the standard multi-head self attention module (MSA) [13]. Each Swin Transformer block is made of layer normalization (ℓ_N), multi-head self-attention and a Multi-Layer Perceptron (MLP) with Gaussian error linear units. The window based self-attention ($\mathrm{W_{sa}}$) and the shifted windows based self- attention ($\mathrm{SW_{sa}}$) modules are applied in consecutive blocks. This process can be summarized in the following operations:

$$\begin{aligned} \widehat{\mathbf{z}}^{l} &= \mathrm{W_{sa}} - \mathrm{MSA}(\ell_N(\mathbf{z}^{l-1})) + \mathbf{z}^{l-1}, \\ \mathbf{z}^{l} &= \mathrm{MLP}(\ell_N(\widehat{\mathbf{z}}^{l})) + \widehat{\mathbf{z}}^{l}, \\ \mathbf{z}^{l+1} &= \mathrm{SW_{sa}} - \mathrm{MSA}(\ell_N(\mathbf{z}^{i})) + \mathbf{z}^{i}, \\ \mathbf{z}^{l+1} &= \mathrm{MLP}(\ell_N(\mathbf{z}^{l+1})) + \mathbf{z}^{l+1}, \end{aligned} \tag{1}$$

where $\widehat{\mathbf{z}}^l$ and $\mathbf{z}^l$ represent the output features of the self-attention modules, computed as in transformer, and the MLP of the l-th block, respectively. On the next $(l+1)$-th block the window attention ($\mathrm{W_{sa}}$) block is replaced by shifted windows ($\mathrm{SW_{sa}}$) and we follow the same procedure as in the previous l-th block (see 2nd and 3rd equations in (1)).

2.2 EfficientNet Encoder

The EfficientNet Encoder [24] based on neural structure search uses composite coefficients to uniformly scale the depth, width, and resolution of the network.

Each EfficientNet block is composed of MBConvBlocks [23] which consists of (*i*) Depthwise convolution, (*ii*) batch normalization, and (*iii*) Swish activation layers. The network achieves better performance with the same parameters by uniformly scaling the network width, depth, or resolution in a fixed proportion. We employ the EfficientNet block as it allows the extraction of several features at different resolutions, which are then used in the skip connections, thus enhancing the model's performance. The architecture used in this case is EfficientNet-B3, which is a scaled up version of EfficientNet-B0. The main blocks of EfficientNet are Convolutional layers and MBconv blocks (see [24]). The MBconv block uses inverted residual connections, making it more efficient [6].

After each EfficientNet Encoder block, the dimensions of the input are reduced by a factor of 2. At each layer before the skip connections, the EfficientNet encoder output is fed to a CLAB block with the corresponding dimensions to that layer (see Fig. 1). The CLAB block provides spatial and channel attention to convolutional models, making it an ideal choice for the EfficientNet branch. It starts by applying a 1×1 convolution to project the input channel dimensions to a set feature number. The resulting features are averaged over the channel dimension which generates spatial attention [14].

2.3 Encoder Fusion and Decoder

Both branches of the encoder are fused at each layer via skip connections and enter the decoder (Fig. 1). However, it is necessary to normalize the dimensions of the outputs of the two encoders since they are of different dimensions. The decoder module consists of four Swin Transformer Decoder blocks that progressively upsample the features extracted by the encoders. At each level, the features resulting from the Swin and the EfficientNet encoders are concatenated together and used via skip connections. The first Swin decoder block employs a patch-expanding layer that upsamples the input features by a factor of 2 in the resolution. The following ones initially project the input to the correct dimensions, then employ a Swin Transformer block, and use the same patch-expanding layer. Finally, the last patch-expanding layer performs a $4\times$ upsampling to restore the resolution of the feature maps to the input size ($\mathrm{W} \times \mathrm{H}$), and a linear projection layer is applied to these up-sampled features to generate pixel-level segmentation predictions.

3 Experimental Setup

We use two datasets, CBIS-DDSM [9] and INbreast [4], to evaluate the performance of our model for the problem of mass lesion segmentation, and compare it with existing methods. Annotations in both datasets include Regions of Interest (ROIs) indicating calcifications and masses, along with Breast Imaging Reporting and Data System (BI-RADS) descriptors, overall BI-RADS assessments, abnormality subtlety ratings, and patient age. INbreast employs full-field digital mammograms, showing better results in cases with dense breasts or microcalcifications. The CBIS-DDSM dataset comprises 1566 cases, with a total of 6775 images, while INbreast consists of 115 cases, making 410 total images. Images are converted to PNG format for ease of use with coding libraries. In both cases, the images are imported in grayscale and resized to 224×224 pixels. We used horizontal flips and color jitter online data augmentation.

The training is conducted on an Intel(R) Core(TM) i7-8700K CPU @ 3.70 GHz, with two graphics cards Nvidia GeForce GTX 1080 Ti 12 Gb and 64 Gb RAM. All models are implemented in Python 3.10.8, Pytorch 1.13.1 and Tensorflow 2.13.0. For both datasets the maximum number of epochs is set to 200. In the Inbreast dataset, the batch size is set to 5 while in CBIS-DDSM it is 16. Both datasets are divided in 80% –20% –20% for training/testing/evaluation. To avoid overfitting, it is implemented early stopping, with a tolerance of 10 epochs, meaning that the training will stop after 10 epochs if there is no improvement in the validation loss. The learning rate is set to $\gamma = 0.0001$.

We compare our proposed model against state-of-the-art baselines. Concretely, we conduct extensive experiments using the following methods for comparison: (*i*) U-Net [7], (*ii*) SETR PUP [10], (*iii*) SwinUNETR [1], (*iv*) TransUNet [5], (*v*) ONET [11], and (*vi*) TFCN [14]. For evaluating the segmentation results we use the following metrics: (*i*) Dice coefficient (*ii*) intersection over union (IoU), and (*iii*) normalized Dice coefficient (nDice) [12].

Metrics (*i*) and (*ii*) are well known and widely used to measure performance. Nonetheless, nDice aims to improve the Dice metric by reducing its bias on the rate of occurrence of the positive class, which tends to favor segmentations where the ground-truth mask is bigger. As such, nDice is particularly relevant for medical image binary segmentation tasks where the target object may only cover a small portion of the entire image.

As in Dice metric, nDice also accesses the similarity between the ground truth and the prediction binary masks by considering the true positives (TP), false positives (FP) and false negatives (FN). Formally, it can be defined as [12]

$$\text{nDice} = 2\left(2 + \rho\frac{\text{FP}}{\text{TP}} + \frac{\text{FN}}{\text{TP}}\right)^{-1}, \tag{2}$$

where $\rho = \ell(\frac{1}{r} - 1)$ with ℓ is the ratio between positive and negative classes in the prediction binary mask and $r \in (0, 1)$ is the occurrence rate of the positive class averaged across all samples in the dataset. Notice that the Dice metric is a particular case of nDice when $\rho = 1$.

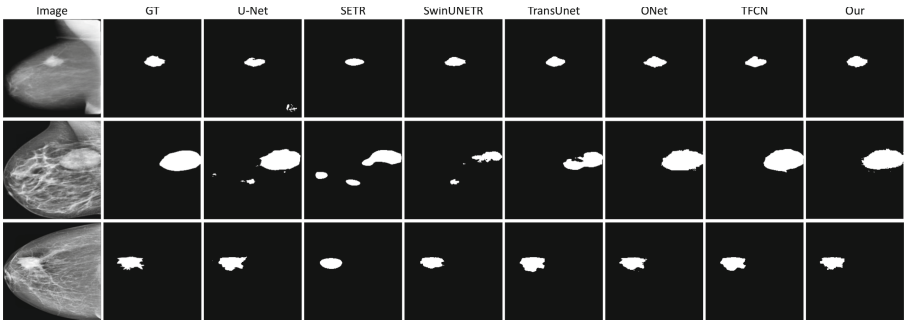


Fig. 2. Segmentations obtained with state-of-the-art methods and our approach

4 Results

This section illustrates the segmentation performance in both CBIS-DDSM and INbreast datasets, as well as a visual analysis of the segmentations obtained and the attention estimated in the CLAB blocks.

4.1 Performance Evaluation

Table 1 summarizes the results for the mass segmentation in both datasets. In Table 1 (a) it is shown that the proposed method performs better in the Dice and nDice metrics, while the baseline ONET outperforms it in the IoU metric. These results show that the addition of the CLAB block to the EfficientNet encoder leads to better model performance. This suggests that the combination of local dependencies extracted from EfficientNet with the global context provided by the CLAB block leads to better feature representations. Similar results were obtained for the INBreast dataset, as shown in Table 1 (b). In this case, the new method outperforms all other methods and on all metrics.

Table 1. Performance evaluation in the CBIS-DDSM (a) and in INbreast (b) datasets.

Model	**Dice**	**nDice**	**IoU**
U-Net [8]	48.1	59.1	38.6
SETR PUP [11]	32.5	42.4	25.5
SwinUNETR [1]	49.1	57.7	41.0
TransUNet [6]	55.8	64.8	44.6
ONET [12]	56.0	67.8	**48.1**
TFCN [15]	54.8	63.8	46.4
Ours	**57.1**	**69.2**	47.5

(a)

Model	**Dice**	**nDice**	**IoU**
U-Net [8]	44.0	53.0	35.2
SETR PUP [11]	31.2	38.0	24.5
SwinUNETR [1]	39.9	44.8	34.8
TransUNet [6]	51.7	55.7	43.0
ONET [12]	58.3	63.6	49.1
TFCN [15]	63.3	70.5	56.5
Ours	**65.9**	**74.7**	**58.8**

(b)

4.2 Visual Analysis

Figure 2 shows examples of segmentations obtained with our approach and other state-of-the-art models. Upon closer examination, it becomes evident that the segmentation produced by our model is closer to the ground truth mask and contains less artifacts than the other state-of-the-art models. Additionally, Fig. 3 (top) shows the evolution of the attention map of the proposed model for a given patient in the INBreast dataset. It first starts by analyzing low level features (*e.g.* edges) and, as the network gets deeper, it shifts the focus to the lesion itself. Figure 3 (bottom) provides more examples using four patients from INBreast dataset, illustrating the last level of attention, where the model is clearly focusing on the lesion in the first three cases. In the fourth case (last example), although the model struggles to focus on the lesion, its attention maps still capture the lesion region. Figure 4 shows more examples using four patients from the CBIS-DDSM dataset (Fig. 4 (a)) and four patients from INBreast dataset (Fig. 4 (b)). We also illustrate the evolution of the attention maps as we go through the levels of the EfficientNet encoder. It is clearly seen the ability of the model at segmenting a large variety of the mass lesion sizes. These correspond to difficult cases as the breast tissue is dense. Note the snapshots in the first row of Fig. 4 (b), where at the $3rd$ level the model is proposing several distinct attention maps in the vicinity of the lesion, but proposing a final saliency map afterwards, correctly highlighting the lesion location.

4.3 Complexity

Hybrid CNN-Transformer models like O-Net typically include approximately 128 million parameters ($\sim$128.3 M), which is considerably higher than standard CNN-based models such as U-Net (which has $\sim$40–70 M parameters). This results in increased memory consumption and inference time. Concretely, we have:

(**1**) **Transformer Branch (Swin Transformer)**: Each self-attention block incurs a complexity of $\mathcal{O}(N \cdot d^2)$, where N is the number of image patches and d is the embedding dimension. This stems from the pairwise interactions in attention. (**2**) **CNN Branch (EfficientNet)**: CNN layers have a complexity of $\mathcal{O}(H \times W \times C \times K^2)$, where H, W are spatial dimensions, C is the number of channels, and K is the kernel size. EfficientNet uses depthwise separable convolutions to reduce cost. Finally, (**3**) **Fusion Layers**: projection layers, skip connection, and decoder fusion layers introduce more computations, but with marginal computational and memory overhead. Compared to pure CNN architectures, our model runs slower due to the attention operations in the transformer branch. Although transformer modules are highly parallelizable, their quadratic complexity makes them less efficient than CNNs for high-resolution inputs.

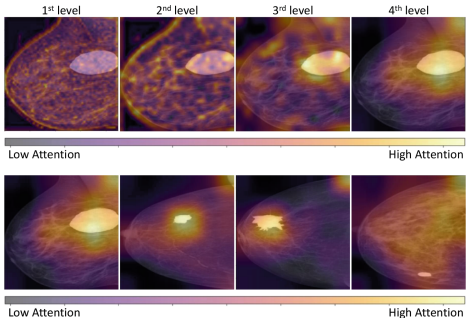
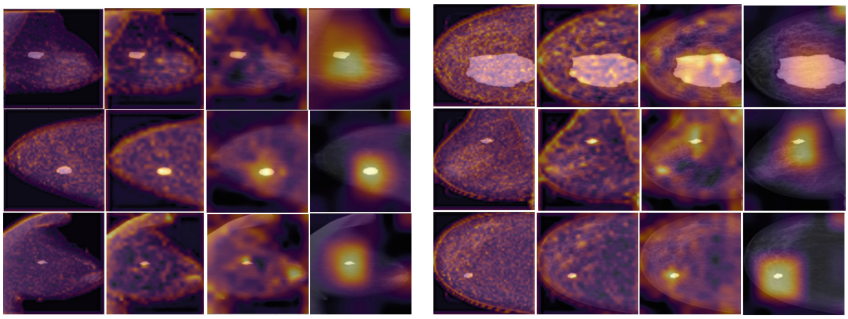


Fig. 3. Attention maps at each level of EfficientNet encoder (top). Attention maps at the last level for different images (bottom)



(a) CBIS-DDSM dataset (b) INBreast dataset

Fig. 4. Illustration of the attention maps evolution for three patients (each patient per row) taken from (a) CBIS-DDSM dataset and (b) INBreast dataset. From the 1*st* to 4*th* columns it is depicted the evolution of the attention maps, as we progress through the levels of the encoder

5 Conclusion

This work proposes a new segmentation model with an encoder that combines an EfficientNet and a Swin Transformer, using a CLAB attention block for feature fusion. The proposed model has the benefits of the local analysis from the CNN-based branch and the long-range context of the Transformer-based branch, while improving representation given to the decoder with spatial and channel attention. Our results show that our approach obtains state-of-the-art results on the segmentation of breast lesions, outperforming all state-of-the-art segmentation models on two breast cancer benchmark datasets.

Acknowledgments. This work was supported by FCT projects MIA-BREAST and AI-Radiologist (10.54499/2022.04485.PTDC and 10.54499/2024.07344.IACDC).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Hatamizadeh, A., et al.: Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images. In: International MICCAI Brainlesion Workshop, pp. 272–284. Springer (2021)
2. Anderson, B.O., et al.: The global breast cancer initiative: a strategic collaboration to strengthen health care for non-communicable diseases. Lancet Oncol. **22**(5), 578–581 (2021)
3. Calisto, F.M., et al.: BreastScreening-AI: evaluating medical intelligent agents for human-AI interactions. Artif. Intell. Med. **127**, 102285 (2022)
4. Moreira, I.C., et al.: Inbreast: toward a full-field digital mammographic database. Acad. Radiol. **19**(2), 236–248 (2012)
5. Chen, J., et al.: TransUNet: transformers make strong encoders for medical image segmentation. CoRR **abs/2102.04306** (2021), https://arxiv.org/abs/2102.04306
6. Howard, A., et al.: Inverted residuals and linear bottlenecks: mobile networks for classification, detection and segmentation. CoRR **abs/1801.04381** (2018), http://arxiv.org/abs/1801.04381
7. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
8. Diogo, P., et al.: Weakly-supervised diagnosis and detection of breast cancer using deep multiple instance learning. In: 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI), pp. 1–4. IEEE (2023)
9. Lee, R.S., et al.: A curated mammography data set for use in computer-aided detection and diagnosis research. Sci. Data **4**(1), 1–9 (2017)
10. Zheng, S., et al.: Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6881–6890 (2021)
11. Wang, T., et al., : O-net: a novel framework with deep fusion of cnn and transformer for simultaneous segmentation and classification. Front. Neurosci. **16** (2022). https://doi.org/10.3389/fnins.2022.876065, https://www.frontiersin.org/articles/10.3389/fnins.2022.876065
12. et al., V.R.: Tackling bias in the dice similarity coefficient: introducing nDSC for white matter lesion segmentation. In: 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI), IEEE (2023)
13. Liu, Z., et al.: Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022 (2021)
14. Li, Z., et al.: TFCNs: a CNN-transformer hybrid network for medical image segmentation. In: Artificial Neural Networks and Machine Learning – ICANN 2022, pp. 781–792. Springer, Cham (2022)
15. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)

16. Dosovitskiy, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
17. Hatamizadeh, A., et al.: Unetr: transformers for 3d medical image segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 574–584 (2022)
18. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
19. Khan, S., Islam, N., Jan, Z., Din, I.U., Rodrigues, J.J.C.: A novel deep learning based framework for the detection and classification of breast cancer using transfer learning. Pattern Recogn. Lett. **125**, 1–6 (2019)
20. Liu, Y., et al.: Roberta: a robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
21. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11976–11986 (2022)
22. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.: Language models are unsupervised multitask learners. OpenAI blog **1**(8), 9 (2019)
23. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv2: inverted residuals and linear bottlenecks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4510–4520 (2018)
24. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning, pp. 6105–6114. PMLR (2019)
25. Valanarasu, J.M.J., Oza, P., Hacihaliloglu, I., Patel, V.M.: Medical transformer: gated axial-attention for medical image segmentation. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12901, pp. 36–46. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87193-2_4
26. Xie, Y., Zhang, J., Shen, C., Xia, Y.: CoTr: efficiently bridging CNN and transformer for 3d medical image segmentation. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12903, pp. 171–180. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87199-4_16
27. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R.R., Le, Q.V.: Xlnet: generalized autoregressive pretraining for language understanding. In: Advances in Neural Information Processing Systems, vol. 32 (2019)



A Domain-Inspired, Semi-supervised nnU-Net Is All You Need for Primary Breast Cancer Segmentation

Daan Schouten(✉), Jeong Hoon Lee, Gregory Bean, Haruka Itakura, and Mirabela Rusu

Stanford University, Stanford 94305, CA, USA
dnschout@stanford.edu

Abstract. An important task in breast cancer detection and prognostication entails the precise delineation of the primary tumor, enabling insight in tumor size, shape characteristics and cancer progression in follow-up studies. However, manual segmentation is labor-intensive and subject to inter-observer variability, motivating the need for accurate automated methods. In light of the ongoing MAMA-MIA challenge, we propose a task-specific and data-centric adaptation of the nnU-Net framework that takes inspiration from the radiologist's workflow for this task. We demonstrate that subtle architectural changes, such as increasing patch size and using multi-channel input, significantly improve performance over a vanilla nnU-Net. Quantitatively, our domain-inspired nnU-Net outperforms a baseline nnU-Net by 3% points in an internal validation setting and 5% points in an external validation setting with additional post-processing methods. Furthermore, we demonstrate that incorporating additional annotated and unannotated data in a semi-supervised manner boosts generalization to unseen domains. In conclusion, our work demonstrates that, despite being a strong general-purpose baseline, a vanilla nnU-Net may still yield a significant performance increase by incorporating domain-specific knowledge. More broadly, this finding corroborates previous research highlighting the need for rigorous validation in medical image segmentation, since common baseline methods may benefit greatly from task-specific adaptation.

Keywords: Breast cancer · Segmentation · Deep learning · Semi-supervised

1 Introduction

Breast cancer remains the most frequently diagnosed malignancy and the leading cause of cancer-related death among women in the United States, with an estimated 314,000 new cases and 43,000 deaths in 2024 alone [11]. Importantly, incidence continues to rise with a 1% annual increase over the 2012–2021 period [3], underscoring the urgent need for improved diagnostic, prognostic, and

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
T. Zhang et al. (Eds.): Deep-Breath 2025, LNCS 16142, pp. 155–164, 2026.
https://doi.org/10.1007/978-3-032-05559-0_16

treatment-monitoring tools. In recent years, dynamic contrast-enhanced magnetic resonance imaging (DCE MRI) has become integral to breast cancer detection and prognostication because of its high sensitivity for detecting malignant lesions, particularly in dense breasts where mammography suffers from reduced sensitivity [12]. An important task in breast cancer diagnosis entails the precise delineation of the primary tumor, enabling insight in tumor shape characteristics (i.e. size or length) and lesion changes over time. Furthermore, accurate delineation of residual tumor during treatment, e.g., chemotherapy, may facilitate monitoring of early response to treatment [8]. However, manual segmentation is labor-intensive and subject to inter-observer variability [4], motivating the need for automated approaches to primary tumor segmentation.

To overcome the limitations of manual segmentation, several deep learning-based methods have recently been proposed for primary tumor segmentation on DCE MRI [2,4,14,15]. Although some methods were shown to perform on par with expert radiologists, validation of these methods was often confined to a modestly sized single-center dataset and generalization to unseen domains often remained underexplored [4,15]. To bridge the gap from single-center to multi-center validation, *Garrucho et al.* recently introduced the MAMA-MIA dataset [2], which aggregates 1506 pre-treatment DCE MRI cases from four The Cancer Imaging Archive (TCIA) collections and contains harmonized expert annotations of the primary lesion. Accompanying this dataset, *Garrucho et al.* introduced the MAMA-MIA challenge, a benchmark for standardized comparison of primary lesion segmentation methods, and trained a vanilla nnU-Net [5] as baseline method for this benchmark. This baseline model obtained a mean dice coefficient of 0.762 in 5-fold cross-validation, demonstrating robustness across multi-institutional data and varying imaging protocols. In extension to these 1506 publicly released cases, the benchmark contains 574 sequestered cases from three centers that were not present in the initial dataset of 1506 cases. Due to the sequestered nature of these external cases, model validation is performed through a containerized code submission to the Codabench platform [13].

In this work, we explore the capabilities of a domain-inspired nnU-Net on this benchmark and investigate the potential of a semi-supervised annotation strategy to further boost performance. Specifically, we investigate the effects of tailoring the nnU-Net architecture for primary breast lesion segmentation and demonstrate that subtle changes may greatly impact performance under identical data conditions. Additionally, we demonstrate that leveraging more annotated and unannotated public data with a semi-supervised annotation strategy yields further improvements in generalization to unseen domains.

2 Methods

Based on the results from the strong nnU-Net baseline released by *Garrucho et al.* and inspired by the clinical task at hand, we make several observations that may further improve the performance of a nnU-Net model in this benchmark setting. In line with previous work [7], we posit that straight-forward task-specific modifications to the nnU-Net framework may be as competitive as novel

methodological approaches. These domain-inspired modifications can be generally divided in data-centric and model-centric, and will be further described in their corresponding subsections.

2.1 Dataset

We primarily rely on the public MAMA-MIA training dataset [2] to develop our method. The MAMA-MIA training dataset bundles previously released public datasets into a single harmonized format, consisting of 1506 DCE MRI cases with expert annotations of the primary lesion. Each primary lesion was determined to be invasive breast cancer based on a targeted biopsy, and DCE MRI scans were acquired before any form of treatment. Due to the multi-center and multi-vendor nature of this dataset, there exists extensive variability in key image characteristics such as the number of phases in the DCE sequence, pixel spacing etc., which have been extensively documented in the original dataset publication [2]. The MAMA-MIA test dataset consists of 574 sequestered cases from three European centers that were not present in the training dataset, making it a truly external validation set. These cases are not publicly available and a performance evaluation on this test dataset can only be achieved through a containerized code submission to the Codabench platform [13].

In addition to the MAMA-MIA training dataset, we leverage additional publicly available annotated and unannotated data to improve data diversity and promote model generalization to unseen domains. These additional datasets can be divided in three main categories, namely annotated invasive cancer cases, unannotated benign lesion cases and unannotated invasive cancer cases. First, we include 63 annotated invasive cancer cases from The Cancer Genome Atlas Breast Invasive Carcinoma Collection (TCGA-BRCA) [9] and 100 annotated invasive cancer cases from the Yunnan Cancer Hospital [14]. Similar to the MAMA-MIA training dataset, all patients in these two additional datasets had biopsy-confirmed invasive cancer and included a primary tumor annotation that was reviewed by an expert radiologist. In light of promoting model generalizability, particularly the dataset from Yunnan Cancer Hospital, originating from Yunnan, China, was deemed to be a valuable addition to the US-centric MAMA-MIA training dataset as the latter did not contain any centers outside of the United States. Second, we include 89 unannotated benign lesion cases and 72 unannotated invasive cancer cases from the Advanced-MRI-Breast-Lesions (AMBL) cohort from The Cancer Imaging Archive (TCIA) [1]. The main uniqueness of the AMBL cohort is that all 161 patients underwent biopsy for a suspicious lesion where histopathological analysis confirmed the benign or malignant nature of the lesion. Hence, the unannotated benign cases can be viewed as difficult negative controls, since the MRI lesion appearance warranted biopsy, but histopathological analysis proved their benign nature. The inclusion strategy of the unannotated invasive cancer cases is further described in the subsection on the semi-supervised annotation strategy. Third, we include a subset of 100 unannotated invasive cancer cases from the Duke-Breast-Cancer-MRI cohort from TCIA [10] that were not present in the MAMA-MIA training dataset. Again, we refer to the

dedicated subsection on the semi-supervised annotation strategy for a further description regarding the handling of these cases. Detailed imaging characteristics for all four additional datasets are captured in Table 1.

Table 1. Main characteristics of the four public datasets that were used in addition to the MAMA-MIA training dataset to train our final model. Values for the number of contrast phases represent median (range) and values for pixel spacing, slice thickness and image size represent the median

	TCGA [9]	Yunnan [14]	AMBL [1]	Duke [10]
N patients	63	100	161	100
N contrast phases	4 (3–6)	5 (5–6)	4 (3–4)	3 (2–4)
Lesion type	Malignant	Malignant	Benign & Malignant	Malignant
Lesion annotation	✓	✓	✗	✗
Pixel spacing	[0.66, 0.66]	[0.38, 0.38]	[0.66, 0.66]	[0.70, 0.70]
Slice thickness	2.0	1.70	2.0	1.0
Image size	[512, 512, 102]	[896, 896, 120]	[512, 512, 116]	[512, 512, 176]

Given the varying DCE MRI protocols that were present in the final combined dataset of 1930 patients, we employ several data pre-processing steps for data harmonization. The breast DCE MRI protocol starts with a single T1-weighted acquisition, the pre-contrast acquisition, followed by a contrast bolus injection and several T1-weighted acquisitions, the post-contrast acquisitions, that aim to visualize the contrast uptake of a potential lesion. Given the variable number of post-contrast acquisitions in this multi-center dataset, we use only the pre-contrast and first post-contrast acquisition. For each patient we resample the pre-contrast and fist post-contrast acquisition to an isotropic $1.0 \times 1.0 \times 1.0$ mm pixel spacing. Second, we employ z-score normalization relative to the pre-contrast acquisition. Specifically, we compute the mean and standard deviation of voxel intensity in the pre-contrast acquisition and use these values to z-score the voxel intensity in the first post-contrast acquisition. This normalization protocol allows for a patient-level normalization that respects intensity peaks related to contrast uptake and better highlights contrast uptake on the subtraction image compared to scan-level z-score normalization.

2.2 Model Architecture

Based on the baseline nnU-Net configuration from *Garrucho et al.*, we propose several domain-inspired adaptions while retaining the original nnU-Net training framework. First, we note that the current benchmark setting revolves around primary lesion segmentation, opposed to segmenting any visible malignant lesion. We therefore hypothesized that the receptive field of the network should cover the entire breast to learn to ignore smaller, secondary lesions in cases with multifocal breast cancer. We thus increase our patch size in the nnU-Net framework

to $192 \times 192 \times 192$, compared to $128 \times 128 \times 128$ in the baseline implementation. Second, we take inspiration from the expert annotation protocol, where radiologists leverage both the first post-contrast phase and the subtraction image (i.e., first post-contrast minus pre-contrast) to annotate the lesion. To provide the model with the same context as the expert annotators, we use two input channels in the nnU-Net framework, namely the first-post contrast scan and the subtraction image. Third, we follow previous research [6] indicating that a larger batch size coupled with a residual encoder may improve training convergence, and increase the batch size to 5 to make use of the full VRAM in the compute node. Lastly, given the benchmark's goal of generalizing to unseen domains, we employ a more aggressive data augmentation strategy with more aggressive spatial transforms and additional contrast-related transforms, thereby emulating unseen MRI domains. The main configuration changes with respect to the baseline are summarized in Table 2. Our final model was trained for 1000 epochs in a 5-fold cross-validation setting on a Linux server with a Nvidia RTX A6000 GPU with 48 GB VRAM, 12 Intel Xeon Gold 5320 CPUs and 512 GB RAM.

Table 2. Architectural differences of the baseline nnU-Net versus our domain-inspired nnU-Net.

	Baseline [2]	Ours
Patch size	$128 \times 128 \times 128$	$192 \times 192 \times 192$
Batch size	2	5
Input channels	1	2
Encoder	Vanilla 3D U-Net	Residual 3D U-net
Data augmentation	Default	Extended

2.3 Semi-supervised Annotation

To leverage the unannotated invasive cancer cases from the AMBL (n = 72) and Duke (n = 100) cohorts in the training process of the final model, we employ a time-efficient semi-supervised annotation strategy to incorporate these cases in our training corpus. We first train our domain-inspired nnU-Net on only the MAMA-MIA training dataset and then use this model to run inference on these unannotated cases in both cohorts. We then manually correct this inferred annotation in a time-efficient manner and feed this case back to the training data for the final model training run. The AMBL cohort originally consisted of 111 unannotated invasive cancer cases, but our initial model did not find any lesion for 39/111 cases, resulting in discarding these 39 cases since annotating the lesion from scratch was infeasible in light of the challenge's time constraints. For the remaining 72/111 cases, we primarily removed false positive secondary lesions from the segmentation mask and corrected major segmentation inaccuracies when present. For the Duke cohort we employed a more targeted annotation

effort, since the original TCIA collection consists of 622 unannotated invasive cancer cases not present in the MAMA-MIA training dataset, which was deemed unfeasible to exhaustively annotate. As a proxy to detect high-quality inferred segmentations, we compute the mean Dice score between all possible combinations of predictions from the five folds of our initial model. In other words, we hypothesize that a high Dice score between predictions from different folds indicates a high inter-model agreement and may serve as a proxy for "easy" cases with a high-quality segmentation. We then sample 30 unifocal and 70 multifocal (which was derived from the cohort's clinical metadata file) breast cancer cases from the total cohort of 622 cases that had the highest inter-model Dice score and manually correct these annotations in a similar fashion to the AMBL cohort. We explicitly sample more multifocal cases since the main failure mode of the initial model was segmentation of additional non-primary lesions (i.e., false positives). Note that, in contrast to the AMBL cohort, the model did not fail to segment any primary lesions in the Duke subset of 100 cases since we explicitly sample cases for which the inter-model Dice coefficient was defined. Our final training dataset (n = 1930) thus encompassed the MAMA MIA training dataset (n = 1506), additional publicly available annotated invasive cancer cases (n = 163), unannotated benign lesion cases (n = 89) and the annotated invasive cancer cases stemming from our semi-supervised annotation strategy (n = 172).

2.4 Post-processing

To further optimize our final model's performance on the external validation set of the MAMA-MIA challenge, we employ two main post-processing steps on the model's output. First, we observe the need to limit the number of cases where the model fails to segment any lesion at all. To combat these failure cases, we employ an iterative segmentation probability thresholding step, where we iteratively decrease our default segmentation probability threshold of 0.50 by 0.05 until we obtain a non-empty segmentation mask. Second, we observe the need to remove additional false positive segmentations in cases with multiple lesion segmentations, since false positive segmentations greatly penalize the normalized Hausdorff distance metric. We therefore propose a false positive reduction method where we first perform a morphological closing operation with a ball-shaped kernel (10 voxel radius) to merge slightly disconnected components from a single lesion, next perform a connected component analysis and ultimately remove any components whose centroid is located further than 50 mm from the largest connected component.

2.5 Statistical Analysis

The performance difference between our proposed method and the baseline method on the 5-fold cross validation set (n = 1506) is assessed using the paired Wilcoxon signed-rank test, since the pairwise performance differences

(proposed method - baseline method) are not normally distributed (Shapiro-Wilk $p<0.001$). We perform a Wilcoxon signed-rank test with one-sided alternative to test whether the median difference exceeds zero, effectively testing superiority of our proposed method relative to the baseline method. We omit statistical testing on the external validation set, as we do not have access to a case-level performance of the baseline method and thus cannot compute pairwise performance differences on this set.

Table 3. Internal and external validation performance comparison of the baseline nnU-Net, the domain-inspired nnU-Net trained on the MAMA-MIA training dataset (n = 1506) and the domain-inspired nnU-Net trained on the extended training dataset (n = 1930). Internal performance values indicate the mean 5-fold cross-validation performance on the MAMA-MIA training dataset (n = 1506) and external performance values indicate the performance on the sequestered validation set (n = 58)

	Baseline [2]		Ours-challenge-data		Ours-extended-data	
	Internal	External	Internal	External	Internal	External
Dice	0.762	0.650	0.795	0.683	0.784	0.701
HD95	37.42	45.00	22.76	18.35	33.14	13.83

3 Results

We test the robustness of our domain-inspired nnU-Net on the original MAMA-MIA training dataset (n = 1506) as well as the external subset that is available in the initial phase of the MAMA-MIA challenge (n = 58). We first examine our initial nnU-Net model, "ours-challenge-data", that was trained under identical data conditions (i.e. exact same folds and number of cases) as the baseline nnU-Net model. The domain-inspired approach resulted in a mean 5-fold cross validation Dice score of 0.795, thereby significantly outperforming the nnU-Net baseline with a Dice score of 0.762 ($p<0.001$). The violin plot of the case-wise Dice scores for both methods in the left subplot of Fig. 1 indicates that this performance improvement primarily stems from a reduction in comparatively low-quality (i.e. Dice < 0.70) segmentations, as the median case-wise performance improvement is relatively modest at +0.0065 Dice. Next, we investigate the effect of our semi-supervised annotation strategy by training another nnU-Net model from scratch on our extended training dataset (n = 1930), which we refer to as "ours-extended-data" in the comparison figure. Note that although this model was trained on additional data in a 5-fold cross-validation setup, we only include the validation metrics on the MAMA-MIA training dataset cases (n = 1506) to enable a fair comparison. Compared to our model trained on only the MAMA-MIA training dataset, we find that our model trained on the extended dataset obtains a slightly reduced mean Dice of 0.784 on the MAMA-MIA training dataset, as this model was not solely optimized for this dataset but rather

a more diverse range of cases. However, the performance increase relative to the baseline nnU-Net remains significant ($p<0.001$) with again a modest median case-wise improvement of +0.0047 Dice. The full internal validation comparison is summarized in Table 3.

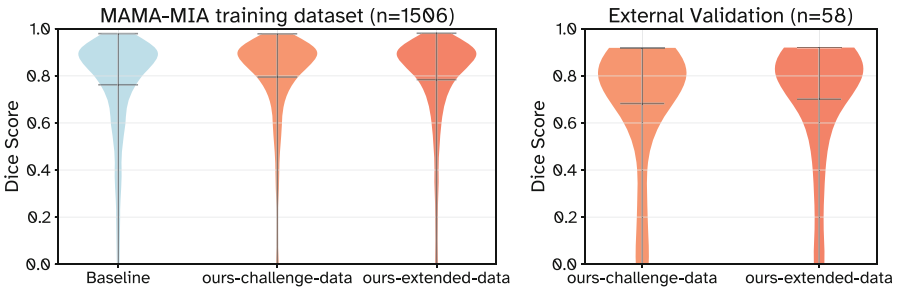


Fig. 1. Comparison of baseline nnU-Net performance and our domain-inspired nnU-Net performance on both the MAMA-MIA training dataset (left) and external validation set (right). Baseline refers to the baseline nnU-Net performance from *Garrucho et al.*, ours-challenge-data refers to our domain-inspired nnU-Net trained on only the MAMA-MIA training dataset (n = 1506) and ours-extended-data refers to our domain-inspired nnU-Net trained on the extended dataset (n = 1930)

The comparison in the external validation scenario is slightly more limited as we do not have access to the case-level Dice score of the baseline model, but only the aggregate Dice score of 0.650 in this setting. Compared to the baseline, our optimized nnU-Net model trained on only the MAMA-MIA training dataset achieves an average Dice score of 0.683 and our model trained on the extended dataset achieves an average Dice score of 0.701, as summarized in Table 3. Although we note that our models incorporate an additional post-processing step relative to the baseline, this step resulted in a modest 0.01 and 0.02 Dice increase, respectively, indicating a persistent improved robustness of our models to unseen domains. The distribution of our model's Dice scores on this external validation set is displayed in the right subplot of Fig. 1. A notable observation of all models in the external setting relates to the generalization gap compared to the MAMA-MIA training dataset, with both a reduction in very high quality segmentations (i.e. Dice score >0.90) and an increase in low quality segmentations (i.e. Dice score <0.50). Importantly, we note that this generalization gap of our models occurred despite an additional post-processing step, which was not employed in the internal validation setting but used in the external validation setting to maximize benchmark performance.

4 Conclusion

We have developed a domain-inspired segmentation model for primary breast cancer segmentation in DCE MRI. We find that, despite being a strong general-purpose baseline, a vanilla nnU-Net may have significant optimization potential

in more sophisticated clinical scenarios (i.e. primary lesion segmentation rather than exhaustive lesion segmentation). By incorporating domain-specific knowledge of the clinical task, we demonstrate a 3.3% point improvement in Dice score in a 5-fold cross-validation setting on an internal validation set. In addition, the optimized nnU-Net trained on an extended dataset with a semi-supervised annotation strategy and a simple post-processing step demonstrates greater generalization capabilities with a 5.0% point improvement in Dice score in an external validation cohort. These findings corroborate an earlier call to action for rigorous benchmarking in medical image segmentation [7], highlighting that simply beating a vanilla nnU-Net baseline with a novel architecture or methodological approach may not necessarily reflect a truly superior approach. Perhaps a more domain- and data-centric approach with an established method such as the nnU-Net framework would have brought about the same sought-after improvement, which may influence time and compute allocation in time-sensitive settings such as biomedical challenges.

Acknowledgments. This work was supported by the National Cancer Institute of the National Institutes of Health under Award Number R37CA260346.

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Daniels, D., Last, D., Cohen, K., Mardor, Y., Sklair-Levy, M.: Standard and delayed contrast-enhanced MRI of malignant and benign breast lesions with histological and clinical supporting data (Advanced-MRI-Breast-Lesions). https://doi.org/10.7937/C7X1-YN57
2. Garrucho, L., et al.: A large-scale multicenter breast cancer DCE-MRI benchmark dataset with expert segmentations. Sci. Data **12**(1), 453 (2025). https://doi.org/10.1038/s41597-025-04707-4
3. Giaquinto, A.N., et al.: Breast cancer statistics 2024. CA: A Can. Jo. Clin. **74**(6), 477–495 (2024), https://doi.org/10.3322/caac.21863, https://onlinelibrary.wiley.com/doi/abs/10.3322/caac.21863, _eprint: https://acsjournals.onlinelibrary.wiley.com/doi/pdf/10.3322/caac.21863
4. Hirsch, L., et al.: Radiologist-level performance by using deep learning for segmentation of breast cancers on MRI scans. Radiol. Artif. Intell. **4**(1), e200231 (2022). https://doi.org/10.1148/ryai.200231, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8823456/
5. Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat. Methods **18**(2), 203–211 (2020)
6. Isensee, F., Maier-Hein, K.H.: Look Ma, no code: fine tuning nnU-Net for the AutoPET II challenge by only adjusting its JSON plans, December 2023. https://doi.org/10.48550/arXiv.2309.13747, http://arxiv.org/abs/2309.13747, arXiv:2309.13747 [eess]
7. Isensee, F., et al.: nnU-net revisited: a call for rigorous validation in 3d medical image segmentation. In: Linguraru, M.G., et al. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2024, pp. 488–498. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-72114-4_47

8. Janse, M.H.A., et al.: Deep learningbased segmentation of locally advanced breast cancer on mri in relation to residual cancer burden: a multiinstitutional cohort study. J. Magn. Reson. Imaging **58**(6), 1739–1749 (2023). https://doi.org/10.1002/jmri.28679
9. Lingle, W., et al.: The cancer genome atlas breast invasive carcinoma collection (TCGA-BRCA). https://doi.org/10.7937/K9/TCIA.2016.AB2NAZRP
10. Saha, A., et al.: Dynamic contrast-enhanced magnetic resonance images of breast cancer patients with tumor locations.https://doi.org/10.7937/TCIA.e3sv-re93
11. Siegel, R.L., Giaquinto, A.N., Jemal, A.: Cancer statistics. CA: A Can. J. Clin. **74**(1), 12–49 (2024). https://doi.org/10.3322/caac.21820, https://onlinelibrary.wiley.com/doi/abs/10.3322/caac.21820, _eprint: https://acsjournals.onlinelibrary.wiley.com/doi/pdf/10.3322/caac.21820
12. Sitges, C., Mann, R.M.: Breast MRI to screen women with extremely dense breasts. J. Magn. Reson. Imaging **62**(1), 58–72 (2025). https://doi.org/10.1002/jmri.29716, https://onlinelibrary.wiley.com/doi/abs/10.1002/jmri.29716, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jmri.29716
13. Xu, Z., et al.: Codabench: flexible, easy-to-use and reproducible benchmarking platform. Patterns **3**(7), 100543 (2022). https://doi.org/10.1016/j.patter.2022.100543, http://arxiv.org/abs/2110.05802, arXiv:2110.05802 [cs]
14. Zhang, J., et al.: A robust and efficient AI assistant for breast tumor segmentation from DCE-MRI via a spatial-temporal framework. Patterns **4**(9), 100826 (2023)
15. Zhang, J., Saha, A., Zhu, Z., Mazurowski, M.A.: Hierarchical convolutional neural networks for segmentation of breast tumors in MRI with application to radiogenomics. IEEE Trans. Med. Imaging **38**(2), 435–447 (2019)

Segmenting Invasive and In Situ Carcinoma in Breast WSIs with a Pretrained Histopathology Transformer

Lauren Jimenez-Martin[1], Felip Vilardell[2], Ana Petit[3], Eduard Dorca[3], Teresa Soler-Monsó[3], Irma Ramos-Oliver[4], Laura Pons[5], Pedro L. Fernández[5], Laia Adalid-Llansa[6], Daniel Mata[7], Lourdes Salazar-Huayna[8], Josep R. Casas[1], and Verónica Vilaplana[1(✉)]

[1] Universitat Politècnica de Catalunya, Barcelona, Spain
{lauren.jimenez,josep.ramon.casas,veronica.vilaplana}@upc.edu
[2] Hospital Universitari Arnau de Vilanova, Lleida, Spain
fvilardell.lleida.ics@gencat.cat
[3] Hospital Universitari de Bellvitge, Hospitalet de Llobregat, Spain
{apetit,edorca,t.soler}@bellvitgehospital.cat
[4] Hospital Universitari de Girona Dr. Josep Trueta, Girona, Spain
iramoso.girona.ics@gencat.cat
[5] Hospital Germans Trias i Pujol, Badalona, Spain
{lponsmar.germanstrias,plfernandez.germanstrias}@gencat.cat
[6] Hospital Universitari Joan XXIII, Tarragona, Spain
ladalidl.hj23.ics@gencat.cat
[7] Hospital de Tortosa Verge de la Cinta, Tortosa, Spain
dmata.ebre.ics@gencat.cat
[8] Hospital Universitari Vall d'Hebron, Barcelona, Spain
lourdes.salazar@vallhebron.cat

Abstract. Accurately distinguishing between invasive carcinoma (IC) and carcinoma in situ (CIS) in breast histopathology is essential for diagnosis and staging, and can inform treatment planning. As digital pathology workflows increasingly integrate AI-based tools, automating this distinction in whole slide images (WSIs) remains a critical challenge due to the subtle morphological differences and the limited availability of detailed annotations.

In this work, we present a multi-class segmentation framework that addresses the segmentation of IC, CIS and benign tissue in hematoxylin and eosin-stained WSIs of core needle biopsies. The model integrates a U-Net decoder with a CTransPath encoder–a transformer-based network pretrained on millions of histopathology image patches–designed to capture both local cellular features and global tissue context. A spatial refinement step is included to reduce misclassification between IC and CIS. The model is fine-tuned on a dataset of 92 WSIs with detailed expert annotations and evaluated at both tile and whole-slide levels. It outperforms baseline models and a non-finetuned version, achieving a macro-averaged tile-level Dice score of 0.80 on the test set. IC is seg-

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
T. Zhang et al. (Eds.): Deep-Breath 2025, LNCS 16142, pp. 165–175, 2026.
https://doi.org/10.1007/978-3-032-05559-0_17

mented with high accuracy (Dice: 0.77), while CIS remains more difficult (Dice: 0.68), reflecting its lower prevalence and more subtle appearance. These results highlight the potential of domain-specific pretrained transformers for robust lesion-level segmentation in breast cancer under real-world constraints, with applications in diagnostic support and case prioritization.

Keywords: breast tumor segmentation · invasive carcinoma · carcinoma in situ · foundation model · transformers

1 Introduction

Breast cancer remains the most frequently diagnosed cancer among women worldwide and a leading cause of cancer-related mortality [3,4]. Accurate histopathological assessment plays a central role in diagnosis, prognosis, and treatment planning. In current clinical practice, pathologists visually examine hematoxylin and eosin (H&E) stained tissue biopsies to characterize tumor extent and type [15]. The digitization of pathology slides into whole slide images (WSIs) has enabled the integration of computational tools into diagnostic workflows. In particular, deep learning approaches have shown considerable promise in WSI analysis, supporting tasks such as tumor region identification, lesion classification, and prognostic feature extraction [11]. However, accurately segmenting distinct histopathological forms of breast carcinoma remains challenging, particularly when attempting to differentiate invasive carcinoma from carcinoma in situ.

Invasive carcinoma (IC) refers to the proliferation of malignant epithelial cells that have breached the basement membrane and infiltrated surrounding tissue [9]. In contrast, carcinoma in situ (CIS) consists of abnormal epithelial cell growth confined to the ductal or lobular spaces, without invasion [2]. This distinction is critical: IC segmentation informs TNM staging [5] and the assessment of biomarkers such as ER, PR, HER2, and Ki-67 [1,13,19]. CIS is also a clinically relevant precursor to IC, and its identification has implications for surveillance and treatment [6].

Despite this clinical importance, most segmentation studies in breast histopathology do not explicitly distinguish CIS from IC. Some approaches group all malignant lesions into a single 'carcinoma' class [7], limiting their diagnostic utility. Others have specifically addressed the IC/CIS distinction using CNN-based models with hierarchical supervision and parallel classification branches [12], but without leveraging recent advances in transformer architectures or domain-specific pretraining. More recently, hybrid CNN–Transformer models have been proposed for IC segmentation, highlighting the advantages of self-attention over convolution for capturing contextual features [9]. However, these models rely on ImageNet-pretrained weights and are trained on patch-level datasets derived from weak or semi-automated annotations. In addition, they are not evaluated on full WSIs with expert-delineated lesion maps, and thus may fall short in supporting clinically meaningful, whole-slide segmentation tasks.

To address this gap, we focus on a diagnostically relevant and technically underexplored task: the explicit segmentation of IC, CIS, and benign tissue in whole-slide images. We propose a segmentation framework that integrates a U-Net decoder with a CTransPath encoder–a transformer-based architecture pretrained on millions of histopathology image patches. Transformers, particularly those pretrained on large-scale histopathology datasets, are well-suited to capturing both local cellular morphology and global contextual relationships. These foundation models learn robust and generalizable representations from unlabeled data, making them highly effective for data-constrained clinical tasks. We finetune our model on a curated dataset of 92 WSIs, fully annotated at the pixel level by expert pathologists from seven institutions. This high-quality annotation setup enables both training and evaluation at the whole-slide level. A spatial refinement step is incorporated to mitigate common misclassifications between IC and CIS. This work is part of an ongoing collaborative effort aimed at supporting deployment in clinical environments. Specifically, the model is designed to assist pathologists by prioritizing WSIs for review and enabling linkage with immunohistochemistry (IHC) slides for further analysis. A comprehensive evaluation at both tile and whole-slide levels demonstrates the potential of this approach for accurate lesion-level segmentation in breast cancer under realistic data constraints.

2 Methodology

2.1 Data

We used a dataset comprising 92 hematoxylin and eosin-stained core needle biopsy whole slide images from different patients, collected from seven hospitals managed by the Catalan Health Institute (ICS), the main public healthcare provider in Catalonia, Spain. A single representative tissue section was annotated per WSI. Annotations were performed independently by two expert pathologists, and any discrepancies were resolved by a third expert, who provided the final ground truth.

The dataset includes annotations for invasive carcinoma and carcinoma in situ. To ensure sufficient representation of each class across evaluation sets, the WSIs were randomly split into training (72), validation (8), and test (12) subsets using a stratified strategy, preserving the proportion of slides containing IC, CIS, or both. Table 1 summarizes the number of WSIs with each tumor category in each split. A notable characteristic of this dataset is class imbalance: in situ lesions are significantly underrepresented compared to invasive carcinoma and benign tissue. Across all slides, 46 contain at least one CIS region, while 86 contain IC, and 40 contain both.

Table 1. Number of WSIs per split containing at least one region of CIS, IC, or both

Split	Total	In Situ	Invasive	Both
Train	72	31	68	27
Validation	8	6	7	5
Test	12	9	11	8
Overall	92	46	86	40

2.2 Preprocessing

To separate tissue from background, we applied a preprocessing pipeline adapted from CLAM [11]. Each WSI was initially downsampled and converted from RGB to HSV color space. We then applied median blurring to smooth edges and binarized the image using a threshold on the saturation channel, which provides better contrast for tissue detection. Morphological closing was used to fill small holes, and area-based filtering removed residual artifacts. This process enabled robust extraction of tissue regions, which were then used to generate tiles for model training. Based on these regions and the expert annotations, we defined a four-class segmentation task: (0) background, (1) benign tissue, (2) carcinoma in situ, and (3) invasive carcinoma.

2.3 Model

Our model is based on the U-Net architecture [14], widely adopted for biomedical image segmentation. We replaced the standard convolutional encoder with CTransPath [18], a Swin Transformer variant [10] optimized for histopathology. CTransPath enhances the Swin Transformer by substituting its patch partition with a Convolutional Neural Network (CNN). This modification leads to more stable training and results in a powerful feature extractor capable of capturing both fine local structures and global context. This encoder was pretrained on millions of unlabeled histopathology patches using their proposed semantically-relevant contrastive learning framework.

By combining U-Net's hierarchical skip connections with the long-range attention of CTransPath, the model captures both fine-grained morphology and global tissue context. The decoder follows the original U-Net design, with successive upsampling and concatenation with encoder features.

The model was trained using a composite loss function combining pixel-wise cross-entropy and Generalized Dice loss [17]. The background class was excluded from loss computation to focus learning on tissue and tumor regions.

2.4 Postprocessing

While not a frequent issue, qualitative inspection revealed that the model occasionally misclassified certain regions of carcinoma in situ as invasive carcinoma,

particularly when CIS formed large contiguous areas and the tile-level appearance lacked clear basement membrane visualization. To mitigate these isolated errors, we implemented a simple spatial postprocessing step.

After reconstructing the whole-slide prediction mask from tile outputs, any connected component labeled as IC was relabeled as CIS if it was entirely or mostly surrounded by CIS regions. This heuristic is based on the observation that such misclassifications tended to occur in enclosed regions where CIS was dominant. This correction was applied uniformly to all test slides but was intended to address only a specific, localized error mode. It did not significantly alter overall performance metrics but improved visual consistency in select edge cases.

3 Experiments and Results

3.1 Training Details

To train and evaluate our segmentation framework, each annotated WSI was divided into patches of 224×224 pixels at 2.5× magnification (2 μm/pixel). Training patches were extracted with 50-pixel overlap to ensure coverage of tissue boundaries and contextual continuity. This process yielded 19,606 training tiles, 1,077 validation tiles, and 1,533 test tiles. Table 2 shows the class distribution as pixel percentages across the training, validation and test sets. As expected, carcinoma in situ (CIS) is underrepresented, accounting for only 2% of pixels in each split, which reflects its lower frequency in biopsy samples.

Table 2. Pixel-wise class distribution (%) across dataset splits

Class	Train	Validation	Test
Background	0.16	0.20	0.15
Tissue	0.43	0.43	0.41
Invasive Carcinoma	0.39	0.36	0.42
Carcinoma In Situ	0.02	0.02	0.02

Model training was performed using the AdamW optimizer with an initial learning rate of 1×10^{-4}, weight decay of 0.1, and a batch size of 64. The model checkpoint with the highest macro-averaged Dice score on the validation set was selected for final evaluation on the test set.

3.2 Results

We evaluated several model configurations to determine the most effective approach for distinguishing invasive and in situ carcinoma. The configurations included: (1) a CTransPath encoder trained from scratch (*Base*), (2) a CTransPath model using pretrained weights without further adaptation

(*CTransPath*), and (3) a CTransPath encoder fine-tuned on our dataset (*Finetuned*). Additionally, we included a 2D adaptation of the *nnUNet* [8,16] framework, which has demonstrated strong performance across various medical imaging tasks.

Table 3 reports tile-level results, including macro-averaged (across the four classes) Dice, precision, and recall, as well as class-specific Dice scores for IC and CIS. The Finetuned model achieved the highest overall macro Dice (0.82 validation, 0.80 test) and the best CIS performance (0.68), while the best IC Dice (0.82) was obtained by the non-finetuned pretrained model. CIS segmentation remained more challenging across models, with consistently lower Dice scores. This likely reflects its lower prevalence in the dataset and the subtle morphological differences that are harder to capture at the tile level, particularly in the absence of clear lesion boundaries or broader tissue context.

Table 3. Tile-level segmentation metrics for all model configurations

Set	Model	Dice	Prec.	Rec.	Dice (IC)	Dice (CIS)
Validation	Base	0.68	0.72	0.67	0.75	0.29
	nnUNet	-	-	-	-	-
	CTransPath	0.79	0.79	0.78	0.82	0.52
	Finetuned	**0.82**	**0.84**	**0.80**	**0.84**	**0.63**
Test	Base	0.69	0.71	0.67	0.75	0.33
	nnUNet	0.77	0.81	0.80	0.68	0.33
	CTranPath	**0.80**	0.81	**0.79**	**0.82**	0.64
	Finetuned	**0.80**	**0.83**	0.78	0.77	**0.68**

WholeSlide Reconstruction: Following tile-based inference, the predicted segmentation maps were reassembled into full-resolution whole-slide images by placing each tile back into its original spatial location. No smoothing, blending, or morphological operations were applied during reconstruction. However, the resulting WSI-level masks were further refined using the spatial postprocessing step described in Sect. 2.4, which reassigns isolated IC predictions surrounded by CIS regions. This allowed for direct visualization and WSI-level assessment of prediction consistency, as illustrated in Fig. 1.

To assess performance at the whole-slide level, we computed precision, recall, and Dice scores for each class independently on each of the 12 test WSIs. Figure 2 shows the per-slide distribution of these metrics for the IC and CIS classes, with each test image represented by a unique color marker. It is important to note that some test slides contain only one of the two tumor classes (IC or CIS), as detailed in Table 1.

Tissue: For this class, the model's performance was generally consistent across WSIs. Both precision and Dice scores typically exceeded 0.8, and recall values

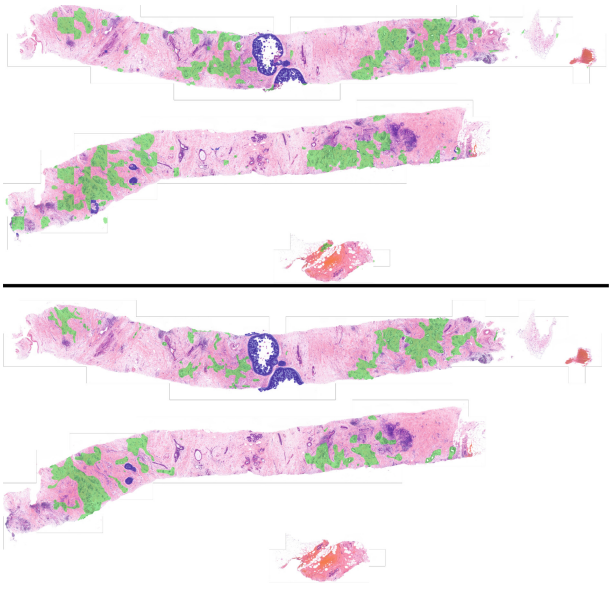


Fig. 1. Model prediction (top) and ground truth annotation (bottom) for tile-level tumor segmentation. IC is shown in green, and CIS in purple. (Color figure online)

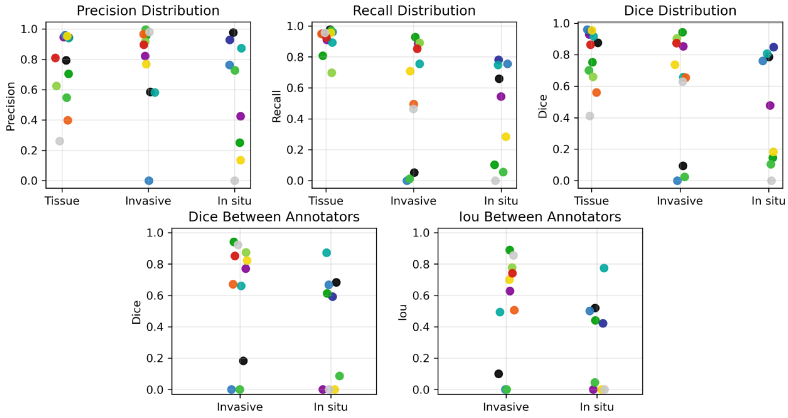


Fig. 2. Top row: Per-class segmentation metrics for benign tissue, IC, and CIS at the whole-slide image level on the test set. Precision, recall, and Dice scores are shown for each test WSI, with uniquely colored points representing individual slides. Bottom row: Inter-annotator agreement in terms of Dice and IoU for IC and CIS annotations across all test slides.

were consistently above 0.9, indicating good agreement with ground truth. One test WSI showed lower precision and Dice, reflecting a specific performance drop for this sample.

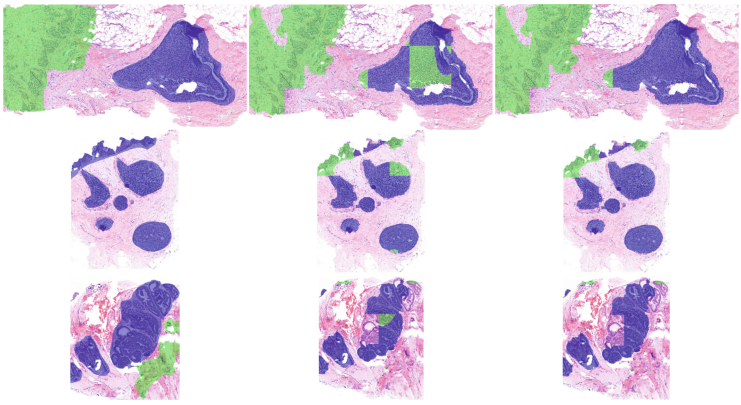


Fig. 3. Visual comparison of segmentation results with post-processing. From left to right: (1) Ground Truth annotations, (2) Initial model predictions, and (3) Post-processed predictions. In all images, IC is denoted in green, and CIS is in purple. (Color figure online)

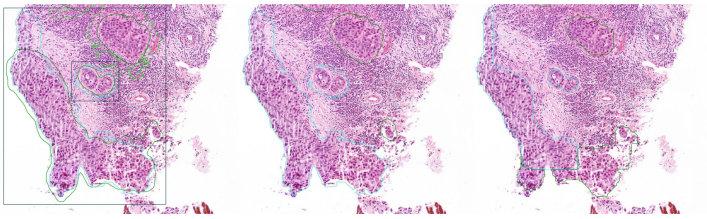


Fig. 4. Left: overlaid annotations from two pathologists (IC in dark/light green, CIS in cyan); middle: final ground truth after review by a third expert; right: model prediction. Discrepancies in two ROIs illustrate inter-observer disagreement and show that the model's output sometimes aligns with one annotator over the final annotation. (Color figure online)

Invasive Carcinoma: The IC class was present in 11 of the 12 test WSIs. For precision, 8 WSIs achieved scores above 0.7. For recall, 6 WSIs scored above 0.7. Dice scores exceeded 0.6 in 8 WSIs. However, 3 distinct cases showed recall and Dice scores near zero. This variation suggests inconsistent segmentation of IC across a subset of WSIs.

Carcinoma in Situ: The CIS class was present in 9 test WSIs and posed notable challenges. For precision, 5 WSIs achieved scores above 0.7. For Dice, 4 WSIs exceeded 0.7. Recall values were more variable: approximately 4 WSIs showed recall near zero. Dice scores were correspondingly low in these cases (below 0.2). The wide distribution of scores reflects both the lower prevalence of CIS and its subtle morphology, which complicates detection. Figure 3 shows representative examples before and after applying the postprocessing heuristic described in Sect. 2.

Annotation Discrepancies: A closer examination of the lowest-scoring CIS and IC cases revealed notable variability in the reference annotations. Figure 2 (bottom row) summarizes the inter-annotator agreement across test slides, showing Dice and IoU between the initial annotations provided by the two pathologists. The mean Dice was 0.60 for IC and 0.39 for CIS, while mean IoU was 0.51 and 0.30, respectively. Notably, for three of the four CIS test WSIs where our model's Dice scores were below 0.2, the overlap between the initial annotations provided by the two pathologists was extremely low (inter-observer Dice below 0.1). In some instances, only one annotator identified a region as tumorous; in others, both did, but with disagreement on whether it corresponded to IC or CIS. A similar pattern was observed for the three lowest-scoring IC cases, where inter-observer agreement was also below 0.2. These discrepancies may partially account for reduced model performance in those slides. Figure 4 illustrates class discrepancies between the initial annotations, the consensus ground truth, and the model prediction on a challenging test WSI, as visualized in QuPath.

4 Conclusion

This study addressed the task of segmenting invasive carcinoma, carcinoma in situ, and benign tissue in hematoxylin and eosin stained core needle biopsy whole-slide images. We proposed a segmentation framework that combines a U-Net decoder with a CTransPath encoder pretrained on large-scale histopathology data and fine-tuned on a curated set of expert-annotated WSIs. Our model achieved strong performance at the tile level, with macro-averaged Dice scores above 0.80 on both validation and test sets. IC and benign tissue were segmented accurately in most cases, while CIS segmentation remained more challenging. Four of nine CIS-containing slides showed low Dice scores (below 0.2) and recall, reflecting the subtle morphology and lower prevalence of CIS in biopsy samples. Analysis of low-performing cases revealed that they often coincided with WSIs where expert annotations also differed, suggesting that model uncertainty arises in slides that are difficult even for human observers. A lightweight postprocessing step improved local consistency in select cases, particularly by correcting small isolated IC predictions within CIS regions. These findings support the potential of histopathology-pretrained transformer encoders for fine-grained tumor segmentation in real-world biopsy samples. Future work will explore strategies for handling annotation uncertainty, incorporating multi-scale and context-aware features, and advancing integration into clinical workflows such as case prioritization and association with IHC analysis.

Acknowledgments. This research was supported by the European Regional Development Fund, under program FEDER Catalunya 2014-2020, SA18-014623 DIGIPATICS; the PID2023-148614OB-I00 research project funded by MICIU/AEI /10.13039/501100011033 and by FEDER, EU and the FI-AGAUR (2022 FI_B 00634) grant funded by Direcció General de Recerca (DGR), Departament de Recerca i Universitats (REU) of the Generalitat de Catalunya and the European Social Fund.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Allison, K.H., et al.: Estrogen and progesterone receptor testing in breast cancer: American society of clinical oncology/college of american pathologists guideline update. Archiv. Pathol. Lab. Med. **144**(5), 545–563 (2020)
2. Allred, D.C.: Ductal carcinoma in situ: terminology, classification, and natural history. J. Natl. Cancer Inst. Monogr. **2010**(41), 134–138 (2010)
3. AM Filho, L.M., Ferlay, J., Colombet, M., Piñeros, M., Znaor, A., et al.: The GLOBOCAN 2022 cancer estimates: data sources, methods, and a snapshot of the cancer burden worldwide. Int. J. Cancer **156**(7), 1336-1346 (2024)
4. Arnold, M., et al.: Current and future burden of breast cancer: global statistics for 2020 and 2040. Breast **66**, 15–23 (2022)
5. Ellis, I., et al.: Dataset for reporting of the invasive carcinoma of the breast: recommendations from the international collaboration on cancer reporting (iccr). Histopathology **85**(3), 418–436 (2024)
6. Farante, G., et al.: Advances and controversies in management of breast ductal carcinoma in situ (DCIS). Eur. J. Surg. Oncol. **48**(4), 736–741 (2022)
7. Ho, D.J., et al.: Deep multi-magnification networks for multi-class breast cancer image segmentation. Comput. Med. Imaging Graph. **88**, 101866 (2021)
8. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat. Meth. **18**(2), 203–211 (2021)
9. Liu, Y., et al.: Invasive carcinoma segmentation in whole slide images using MS-ResMTUNet. Heliyon **10**(4) (2024)
10. Liu, Z., et al.: Swin Transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022 (2021)
11. Lu, M.Y., et al.: Data-efficient and weakly supervised computational pathology on whole-slide images. Nat. Biomed. Eng. **5**(6), 555–570 (2021)
12. Ni, H., Liu, H., Wang, K., Wang, X., Zhou, X., Qian, Y.: WSI-Net: branch-based and hierarchy-aware network for segmentation and classification of breast histopathological whole-slide images. In: Machine Learning in Medical Imaging: 10th International Workshop, MLMI 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 13, 2019, Proceedings 10, pp. 36–44. Springer (2019)
13. Nielsen, T.O., et al.: Assessment of Ki67 in breast cancer: updated recommendations from the international Ki67 in breast cancer working group. JNCI: J. Nat. Cancer Inst. **113**(7), 808–819 (2021)
14. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional Networks for Biomedical Image Segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
15. Sharma, G.N., Dave, R., Sanadya, J., Sharma, P., Sharma, K.: Various types and management of breast cancer: an overview. J. Adv. Pharm. Technol. Res. **1**(2), 109–126 (2010)

16. Spronck, J., et al.: nnUNet meets pathology: bridging the gap for application to whole-slide images and computational biomarkers. In: Medical Imaging with Deep Learning (2023)
17. Sudre, C.H., Li, W., Vercauteren, T., Ourselin, S., Jorge Cardoso, M.: Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, Proceedings 3, pp. 240–248. Springer (2017)
18. Wang, X., et al.: Transformer-based unsupervised contrastive learning for histopathological image classification. Med. Image Anal. **81**, 102559 (2022)
19. Wolff, A.C., et al.: Human epidermal growth factor receptor 2 testing in breast cancer: American society of clinical oncology-college of American pathologists guideline update. Archiv. Pathol. Lab. Med. **147**(9), 993–1000 (2023)



SiameseOrdinalCLIP: A Language-Guided Siamese Network for the Aesthetic Evaluation of Breast Cancer Locoregional Treatment

Fábio Teixeira[1(✉)], Helena Montenegro[1], Eduard Bonci[2], Maria J. Cardoso[2,3], and Jaime S. Cardoso[1]

[1] INESC TEC, Faculdade de Engenharia, Universidade do Porto, Rua Dr. Roberto Frias, 4200-465 Porto, Portugal
fabioat09@gmail.com
[2] Fundação Champalimaud, Av. Brasília, 1400-038 Lisbon, Portugal
[3] Faculdade de Medicina, Universidade de Lisboa, Av. Prof. Egas Moniz MB, 1649-028 Lisbon, Portugal

Abstract. Breast cancer locoregional treatment includes a wide variety of procedures with diverse aesthetic outcomes. The aesthetic assessment of such procedures is typically subjective, hindering the fair comparison between their outcomes, and consequently restricting evidence-based improvements. Most objective evaluation tools were developed for conservative surgery, focusing on asymmetries while ignoring other relevant traits. To overcome these limitations, we propose SiameseOrdinalCLIP, an ordinal classification network based on image-text matching and pairwise ranking optimisation for the aesthetic evaluation of breast cancer treatment. Furthermore, we integrate a concept bottleneck module into the network for increased explainability. Experiments on a private dataset show that the proposed model surpasses the state-of-the-art aesthetic evaluation and ordinal classification networks.

Keywords: Aesthetic Evaluation · Breast Cancer Treatment · Medical Imaging · Ordinal Classification

1 Introduction

With the diversification of treatment and improved screening, breast cancer has become an increasingly treatable disease with a high survival rate [3]. Locoregional treatment options include a wide range of surgeries focusing on the entire (mastectomy) or partial (conservative treatment) removal of the affected breast(s), optionally followed by reconstructive procedures. These treatments can have diverse aesthetic impacts on a patient's breasts, often causing asymmetries, visible scars and other deformations. The high diversity of surgical treatments with similar oncological outcomes has led to an increased need for methods to evaluate and compare the aesthetic impact of different procedures.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
T. Zhang et al. (Eds.): Deep-Breath 2025, LNCS 16142, pp. 176–185, 2026.
https://doi.org/10.1007/978-3-032-05559-0_18

The aesthetic assessment of breast cancer treatment has relied mostly on patient-reported metrics [19] and panels of experts [17,23]. However, these assessments are not reproducible nor reliable due to the patients' lack of knowledge regarding potential outcomes of treatment and the inherent subjectivity of the evaluation task [5]. As a result, subjective evaluation should not be used as a gold standard to compare aesthetic outcomes. Recently, various tools have been proposed for objective evaluation, ranging from objective asymmetry measurements [16,22] to machine learning tools that predict a score representative of the aesthetic outcome [3]. Still, most methods were developed specifically for conservative treatment, focusing mostly on the asymmetries between the breasts. Furthermore, most machine learning methods require human intervention and do not consider the ordinal relation between the aesthetic scores.

Various ordinal classification networks, that explicitly model the ordinal relationship between existing classes, have been proposed in the scientific literature. For example, R^2-ResNeXt [14] trains a Siamese network to model the relative ranking of pairs of images for face beauty prediction. More recently, OrdinalCLIP [13] aligns image embeddings with class embeddings in an ordinal space learned by a natural language model. Nevertheless, these works have not been applied to the aesthetic assessment of breast cancer treatment.

In this work, we propose SiameseOrdinalCLIP, a fully automatic ordinal classification network for the aesthetic evaluation of breast cancer treatment. SiameseOrdinalCLIP uses the image encoder of OrdinalCLIP [13] as a backbone to a Siamese network that simultaneously optimises the prediction of the aesthetic score and the pairwise aesthetic ranking between breast images, inspired by R^2-ResNeXt [14]. Furthermore, we perform an experiment showing how concepts such as asymmetry measurements can be integrated into the network for increased explainability. Experiments on a private dataset show that our network surpasses state-of-the-art ordinal classification and aesthetic evaluation models.

2 Related Work

This section provides an overview of the existing work on the aesthetic evaluation of breast cancer treatment and ordinal classification.

2.1 Aesthetic Evaluation of Breast Cancer Treatment

Traditional works on the aesthetic evaluation of breast cancer treatment measure asymmetries between the breasts. More specifically, various methods have been proposed to evaluate the displacement of the nipples [16,21,22], and differences between the size, shape and colour of the breasts [3,22]. Nevertheless, these strategies require manual annotation of breast contour and are sensitive to the patient's posture. More recently, various strategies incorporate asymmetry measurements to assign an aesthetic score to an image. For example, BCCT.core [3] uses a Support Vector Machine for the final prediction. Guo et al. [9] experiment with various machine learning models to map objective metrics into an aesthetic

score. Silva et al. [20] condition a deep learning model on asymmetry metrics. All these works were developed for conservative treatment, where asymmetries are prominent, and have limited applicability in oncoplastic surgeries with smaller asymmetries and more visible scars. Furthermore, despite the ordinal nature of the evaluation task, the ordinal relationship between the aesthetic classes remains largely underexplored in the current literature.

2.2 Ordinal Classification

Ordinal classification can be achieved through an ensemble of binary classifiers [6,8]. Nevertheless, this strategy can obtain contradictory results. Other strategies include constraining the output of the network to be unimodal (e.g., by modelling the task as a regression) [2,7], or promoting ordinality through a loss function [1,4]. Recent works focus on improving the ordinality of the latent space of deep learning networks. For example, OrdinalCLIP [13] leverages the latent space of a pretrained language model (CLIP [18]) to learn an ordinal set of embeddings representative of each class, using a contrastive loss to approximate the embeddings of an image to the corresponding class. R^2-ResNeXt [14] uses a Siamese network to optimize the pairwise ranking of images within the latent space of a ResNeXt [24] optimized for regression.

3 Methodology

Our goal is to map a post-surgical anterior photograph to an ordinal label $y \in \{1:\textit{poor}, 2:\textit{fair}, 3:\textit{good}, 4:\textit{excellent}\}$, according to the Harvard scale [10]. To achieve this goal, we develop a deep neural network, SiameseOrdinalCLIP, which leverages ordinal information to guide the classification process. SiameseOrdinalCLIP's training proceeds in two stages, described in the following sections.

3.1 Stage 1: Language-Guided Ordinal Pre-Training

In the first stage, we use OrdinalCLIP [13] to obtain an image encoder that learns representations of breast images aligned with the corresponding class in an ordinal latent space.

OrdinalCLIP trains an image encoder to align image embeddings with language prototypes $\{\mathbf{p}_c\}_{c=1}^{4}$ produced by a frozen text encoder from CLIP [18], as illustrated in Fig. 1. The language prototypes are obtained by feeding a prompt into the text encoder. Nevertheless, since different prompts can have different results, these tokens are trainable, and OrdinalCLIP learns the prompt that optimises the ordinal classification task. To facilitate learning, we initially instantiate the prompt as

$$\text{“}\texttt{Aesthetic assessment: As a professional surgeon,}\ \texttt{I would rate the aesthetic outcome of this breast surgery as ''} \quad (1)$$

This prompt is then concatenated with the image's label before being fed to the text encoder. Since the labels *poor/fair* are ambiguous and convey little semantic separation to a language model, we replace them with the harsher distance-amplifying adjectives *"horrible"* and *"bad"*, respectively.

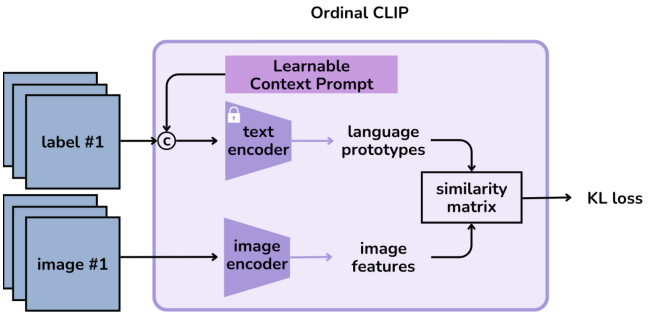


Fig. 1. Diagram for training stage 1

After the image encoder (ResNet-50 [11]) and the prompt tokens are jointly optimised, using the bi-directional contrastive loss $\mathcal{L}_{\text{CLIP}}$ from the original paper [13], we discard the text branch and keep the image encoder.

3.2 Stage 2: Siamese Ordinal Fine-Tuning

In the second stage, the frozen image encoder backbone obtained from Stage 1 feeds a smaller network inspired by *R2-ResNeXt* [14] and illustrated in Fig. 2, composed of two lightweight heads that share a common 2-layer multilayer perceptron (MLP):

- **Regression head.** A linear layer outputs $z \in \mathbb{R}$, rescaled to the continuous interval $[0.5, 4.5]$ via $\hat{y} = 0.5 + 4\sigma(z)$, ensuring equal dynamic range for all classes, where σ represents the sigmoid function.
- **Ranking head.** Given a pair of images $(\mathbf{x}_i, \mathbf{x}_j)$, the concatenated embedding $[\mathbf{f}_i \| \mathbf{f}_j]$ go through a linear layer that predicts a signed score s_{ij} whose sign indicates which of the images has a better aesthetic score.

Training objective. For a mini-batch $\mathcal{B}$ with a set of annotator-consistent, label-discordant pairs $\mathcal{P}$, we minimize the mean squared error ($\mathcal{L}_{MSE}$) and hinge loss ($\mathcal{L}_{hinge}$) according to the following equation:

$$\mathcal{L} = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \underbrace{(\hat{y}_i - y_i)^2}_{\mathcal{L}_{\text{MSE}}} + \lambda \left[\frac{1}{|\mathcal{P}|} \sum_{(i,j) \in \mathcal{P}} \underbrace{\max(0,\, \alpha_{ij} - \eta_{ij} s_{ij})}_{\mathcal{L}_{\text{hinge}}} \right], \tag{2}$$

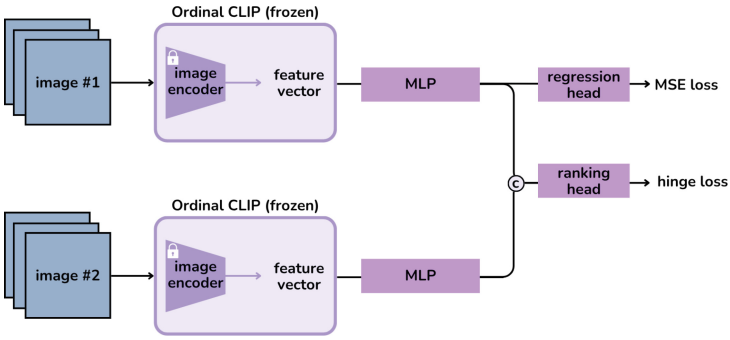


Fig. 2. Diagram with final network

where $\eta_{ij} = +1$ if $y_i > y_j$, -1 otherwise, and the adaptive margin $\alpha_{ij} = |y_i - y_j|/2$ increases the penalty for widely separated ranks. We set $\lambda = 0.5$ as a parameter to calibrate the relative importance of the hinge loss.

This Siamese portion of the network improves on the ordinality learned in the previous stage by directly comparing the aesthetic outcomes of different breast images using the ranking head. For inference, the aesthetic class assigned to an image is the nearest integer to the output of the regression head ($\hat{y}$).

3.3 Data Augmentation Strategies

To overcome the relatively low amount of data and balance the less represented classes (*poor* and *excellent*), different data augmentation techniques were used:

- **Breast Manipulation** - Synthetic images were generated by feeding post-operative images into the breast manipulation algorithm proposed by Montenegro et al. [15], exacerbating negative visual characteristics, such as asymmetry. These images were used to supplement the *poor* class (Fig. 3).
- **Pre-operative Images** - Pre-operative images were hand-selected and used to supplement the *excellent* class (Fig. 4).

In addition to these strategies, we perform traditional data augmentation through horizontal flips, $\pm 15°$ rotations and mild colour jitter to augment all classes during training.

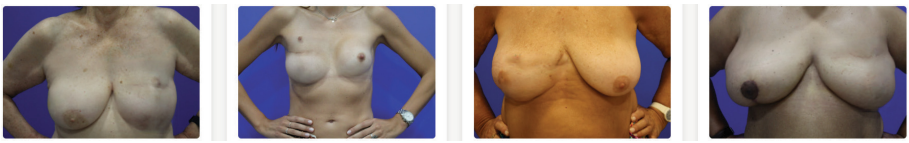


Fig. 3. Examples of synthetic images generated through morphing to supplement the *poor* class

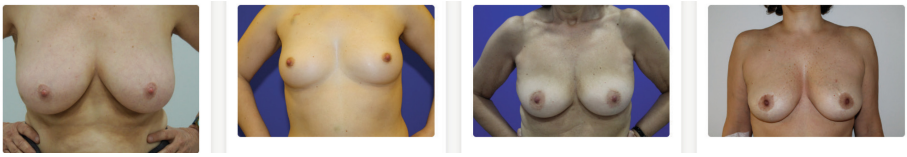


Fig. 4. Examples of handpicked preoperative images to supplement the *excellent* class

3.4 Explainability

Finally, we integrate *concept bottleneck* (CB) modules into the SiameseOrdinalCLIP model to force the network to predict explicit human-readable attributes (e.g. breast symmetry), before predicting the final score, as shown in Fig. 5. By integrating asymmetry measurements (particularly relevant for conservative treatment) into the model, we obtain a more interpretable model that relies on human-understandable concepts for the prediction and whose predictions can be used as explanations to understand the reasoning of the model.

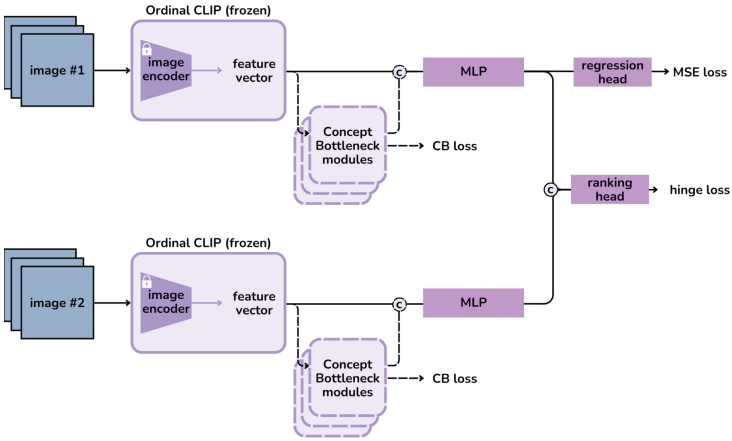


Fig. 5. Final network with CB modules

Each module is a small MLP with sigmoid activation that outputs a scalar in the [0, 1] range representing a clinical concept. The CB modules are tuned in the second training stage using a mean squared error (MSE) loss and Lower Breast Contour [22] measurements as symmetry labels. The new training objective for stage two can be described as:

$$\mathcal{L} = \underbrace{\mathcal{L}_{\text{MSE}}(\hat{y}, y)}_{\text{regression}} + \lambda \underbrace{\mathcal{L}_{\text{hinge}}(s_{ij})}_{\text{ranking}} + \gamma \underbrace{\mathcal{L}_{\text{MSE}}\big(c_{\text{sym}}, t_{\text{sym}}\big)}_{\text{concept}},$$

where t_{sym} is the ground-truth symmetry score derived from LBC measurements and $\lambda = 0.5$ and $\gamma = 0.1$ are used to calibrate the relative importance of

each loss term. Finally, the predicted asymmetry values are concatenated with the image's features and provided to the regression head for the final prediction.

4 Experiments and Results

The experiments use 1,406 anonymised frontal photographs of postoperative breasts from a private dataset obtained from the portuguese Champalimaud Foundation in the context of the CINDERELLA project [12]. Each image is labelled by one of three senior surgeons according to the Harvard scale [10]. The dataset is imbalanced, with the majority of images belonging to the fair (26%) and good (43%) classes. The poor and excellent classes are only represented in 8% and 23% of the images, respectively. Regarding the type of surgery, 54% of the postoperative images represent conservative surgery without reconstruction, while the remaining 46% represent various oncoplastic surgeries. We divided the data into 70% for training, 15% for validation and 15% for testing. These splits are consistent in the two training stages of SiameseOrdinalCLIP.

In the experiments, we compare the results of the proposed SiameseOrdinalCLIP with four baselines and state-of-the-art works:

- **ResNet-50** [11]: ResNet-50 fine-tuned using MSE loss.
- **BCCT.core** [3]: SVM classification based on asymmetry measurements.
- **OrdinalCLIP** [13]: our Stage 1 backbone with a single regression head.
- **R2-ResNeXt** [14]: our Stage 2 network with a ResNeXt-50 backbone.

Additionally, we perform various ablation studies to:

- Measure the impact of the alteration of the fair/poor classes to bad/horrible on OrdinalCLIP.
- Measure the impact of the pairwise ranking loss by altering λ on R^2-ResNext.
- Measure the impact of the augmentation of the poor and excellent classes (+ Aug) on the proposed SiameseOrdinalCLIP.
- Measure the impact of the concept bottleneck module (+ CB) on the proposed network.

Table 1 summarises the performance of all the models. Since the task is ordinal and class-imbalanced, we report (i) Root Mean Squared Error (RMSE), (ii) Mean Average Error (MAE), (iii) Macro-Averaged MAE (M-MAE), (iv) Accuracy (Acc), and (v) Ranking Accuracy (Rank-Acc).

As shown by the results, SiameseOrdinalCLIP with data augmentation surpasses the state-of-the-art aesthetic evaluation model BCCT.core in all metrics, with an increase of over 8% in accuracy and over 30% in ranking accuracy, showing the classification process is much more consistent with the experts' opinions. Furthermore, the proposed network surpasses the state-of-the-art ordinal classification networks, suggesting that the ordinal space learned by OrdinalCLIP benefits from further optimisation of the pairwise ranking between images. In the experiments with R^2-ResNeXt, the configuration with $\lambda = 0.5$ obtained slightly

Table 1. Aesthetic evaluation results

Model	Acc ↑	Rank-Acc ↑	RMSE ↓	MAE ↓	M-MAE ↓
BCCT.core [3]	0.5721	0.5649	0.7153	0.4558	0.6057
ResNet-50 [11]	0.4834	0.7635	0.7994	0.6246	0.6029
R^2-ResNeXt [14] $\lambda = 0.1$	0.4787	0.6890	0.8254	0.6611	0.7600
R^2-ResNeXt [14] $\lambda = 0.5$	0.5071	0.6963	0.8137	0.6527	0.7317
R^2-ResNeXt [14] $\lambda = 0.9$	0.4976	0.6887	0.8183	0.6529	0.7471
OrdinalCLIP [13]: fair/poor	0.4834	0.4902	0.8290	0.5735	0.6852
OrdinalCLIP [13]: bad/horrible	0.5166	0.5360	0.7910	0.5308	0.6017
SiameseOrdinalCLIP	0.6114	0.8616	0.6300	0.4856	0.3394
SiameseOrdinalCLIP + Aug	**0.6588**	**0.8773**	**0.5641**	**0.4510**	**0.3104**
SiameseOrdinalCLIP + Aug + CB	0.6161	0.8463	0.6142	0.4974	0.3838

better results than the remaining configurations. The experiments with OrdinalCLIP suggest that replacing the labels *fair* and *poor* by *bad* and *horrible*, respectively, improves the results. Finally, the addition of the CB module into the proposed SiameseOrdinalCLIP introduces explainability into the network at the cost of a slight decrease in performance.

To assess the explainability module, we manually alter the predicted symmetry score of the images in the test set to 0 (fully asymmetric) and 1 (perfectly symmetric) at inference time, and measure the alteration to the prediction (Δ_0 and Δ_1, respectively). The model's output shifted by an average of $\Delta_0 = -0.618$ and $\Delta_1 = +1.833$ points, demonstrating a strong relationship between symmetry and the aesthetic score. Figure 6 shows an example of an explanation, highlighting the impact of the symmetry alteration on a test image.

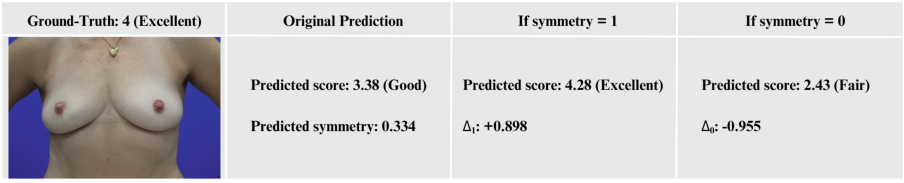


Fig. 6. Example of explanation produced by SiameseOrdinalClip

5 Conclusions

We introduced **SiameseOrdinalCLIP**, the first fully automatic, language-guided ordinal classifier for aesthetic evaluation across diverse breast cancer surgeries. By coupling CLIP-based ordinal pre-training with a lightweight Siamese

ranking head, the model is able to perceive ordinality to a much higher degree than base architectures, better matching the experts' decision process, and achieving better results over the gold-standard BCCT.core software, while not requiring any manual intervention. Future work will focus on validating the architecture on external datasets, improving explainability, increasing data availability and experimenting with providing a wider range of clinical (e.g., type of surgery and more asymmetry measurements) and demographic (e.g., cup size) concepts to the network through the concept bottleneck modules.

Acknowledgments. This work has received funding from the European Union's Horizon Europe research and innovation programme under the Grant Agreement 101057389-CINDERELLA project, and from the FCT - Foundation for Science and Technology Portugal, within PhD grant 2022.14516.BD.

Disclosure of Interests. The authors have no competing interests to declareăthat are relevant to the content of this article.

References

1. Albuquerque, T., Cruz, R., Cardoso, J.S.: Ordinal losses for classification of cervical cancer risk. PeerJ Comput. Sci. **7**, e457 (2021)
2. Beckham, C., Pal, C.: Unimodal probability distributions for deep ordinal classification. In: International Conference on Machine Learning, pp. 411–419. PMLR (2017)
3. Cardoso, J.S., Cardoso, M.J.: Towards an intelligent medical system for the aesthetic evaluation of breast cancer conservative treatment. Artif. Intell. Med. **40**(2), 115–126 (2007)
4. Cardoso, J.S., Cruz, R.P.M., Albuquerque, T.: Unimodal distributions for ordinal regression. IEEE Trans. Artif. Intell. 1–12 (2025). https://doi.org/10.1109/TAI.2025.3549740
5. Cardoso, J.S., Silva, W., Cardoso, M.J.: Evolution, current challenges, and future possibilities in the objective assessment of aesthetic outcome of breast cancer locoregional treatment. Breast **49**, 123–130 (2020)
6. Cheng, J., Wang, Z., Pollastri, G.: A neural network approach to ordinal regression. In: 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), pp. 1279–1284. IEEE (2008)
7. da Costa, J.P., Cardoso, J.S.: Classification of ordinal data using neural networks. In: Gama, J., Camacho, R., Brazdil, P.B., Jorge, A.M., Torgo, L. (eds.) ECML 2005. LNCS (LNAI), vol. 3720, pp. 690–697. Springer, Heidelberg (2005). https://doi.org/10.1007/11564096_70
8. Frank, E., Hall, M.: A simple approach to ordinal classification. In: De Raedt, L., Flach, P. (eds.) ECML 2001. LNCS (LNAI), vol. 2167, pp. 145–156. Springer, Heidelberg (2001). https://doi.org/10.1007/3-540-44795-4_13
9. Guo, C., et al.: A fully automatic framework for evaluating cosmetic results of breast conserving therapy. Mach. Learn. Appl. **10**, 100430 (2022)
10. Harris, J.R., Levene, M.B., Svensson, G., Hellman, S.: Analysis of cosmetic results following primary radiation therapy for stages i and ii carcinoma of the breast. Int. J. Radiat. Oncol. Biol. Phys. **5**(2), 257–261 (1979). https://doi.org/10.1016/0360-3016(79)90729-6

11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
12. Kaidar-Person, O., et al: Evaluating the ability of an artificial-intelligence cloud-based platform designed to provide information prior to locoregional therapy for breast cancer in improving patient's satisfaction with therapy: the CINDERELLA trial .https://doi.org/10.1371/journal.pone.0289365, https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0289365
13. Li, W., et al.: Ordinalclip: learning rank prompts for language-guided ordinal regression. Adv. Neural. Inf. Process. Syst. **35**, 35313–35325 (2022)
14. Lin, L., Liang, L., Jin, L.: R 2-resnext: A resnext-based regression model with relative ranking for facial beauty prediction. In: 2018 24th International Conference on Pattern Recognition (ICPR), pp. 85–90. IEEE (2018)
15. Montenegro, H., Cardoso, M.J., Cardoso, J.S.: An inpainting approach to manipulate asymmetry in pre-operative breast images. arXiv preprint arXiv:2502.05652 (2025)
16. Pezner, R.D., et al.: Breast retraction assessment: an objective evaluation of cosmetic results of patients treated conservatively for breast cancer. Int. J. Radiat. Oncol. **11**(3), 575–578 (1985)
17. Pierquin, B., et al.: Conservative treatment for breast cancer: long-term results (15 years). Radiother. Oncol. **20**(1), 16–23 (1991)
18. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748–8763. PmLR (2021)
19. Seth, I., Seth, N., Bulloch, G., Rozen, W.M., Hunter-Smith, D.J.: Systematic review of breast-q: a tool to evaluate post-mastectomy breast reconstruction. Breast Can. Targets Ther. 711–724 (2021)
20. Silva, W., Carvalho, M., Mavioso, C., Cardoso, M.J., Cardoso, J.S.: Deep aesthetic assessment and retrieval of breast cancer treatment outcomes. In: Iberian Conference on Pattern Recognition and Image Analysis, pp. 108–118. Springer (2022)
21. Tsouskas, L., Fentiman, I.: Breast compliance: a new method for evaluation of cosmetic outcome after conservative treatment of early breast cancer. Breast Cancer Res. Treat. **15**, 185–190 (1990)
22. Van Limbergen, E., Van der Schueren, E., Van Tongelen, K.: Cosmetic evaluation of breast conserving treatment for mammary cancer. 1. proposal of a quantitative scoring system. Radiother. Oncol. **16**(3), 159–167 (1989)
23. Xia, T.Y., Scomacao, I., Duraes, E., Cakmakoglu, C., Schwarz, G.: Aesthetic, quality-of-life, and clinical outcomes after inferior pedicle oncoplastic reduction mammoplasty. Aesthetic Plast. Surg. **47**(3), 905–911 (2023)
24. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1492–1500 (2017)



Anatomically and Clinically Informed Deep Generative Model for Breast Surgery Outcome Prediction

Joana Santos[1(✉)], Helena Montenegro[1], Eduard Bonci[2], Maria J. Cardoso[2,3], and Jaime S. Cardoso[1]

[1] INESC TEC, Faculdade de Engenharia, Universidade do Porto, Rua Dr. Roberto Frias, 4200-465 Porto, Portugal
joana.i.santos@hotmail.com
[2] Fundação Champalimaud, Av. Brasília, 1400-038 Lisbon, Portugal
[3] Faculdade de Medicina, Universidade de Lisboa, Av. Prof. Egas Moniz MB, 1649-028 Lisbon, Portugal

Abstract. Breast cancer patients often face difficulties when choosing among diverse surgeries. To aid patients, this paper proposes ACID-GAN (Anatomically and Clinically Informed Deep Generative Adversarial Network), a conditional generative model for predicting post-operative breast cancer outcomes using deep learning. Built on Pix2Pix, the model incorporates clinical metadata, such as surgery type and cancer laterality, by introducing a dedicated encoder for semantic supervision. Further improvements include colour preservation and anatomically informed losses, as well as clinical supervision via segmentation and classification modules. Experiments on a private dataset demonstrate that the model produces realistic, context-aware predictions. The results demonstrate that the model presents a meaningful trade-off between generating precise, anatomically defined results and maintaining patient-specific appearance, such as skin tone and shape.

Keywords: Breast Cancer · Deep Generative Models · Image-to-Image Translation · Post-operative Outcome Prediction · Medical Image Synthesis

1 Introduction

Breast cancer treatments have become increasingly effective, with modern protocols often combining surgery, radiotherapy, and systemic therapies [1,8]. Among surgical options, patients may choose between breast-conserving surgery and mastectomy with or without reconstruction, procedures that, while comparable in oncological outcomes, can result in very different aesthetic changes. Communicating these differences is challenging, especially when patients must form expectations about highly individualised outcomes like symmetry, scarring, and skin tone [9]. Current tools, such as diagrams or reference photos, are limited in

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
T. Zhang et al. (Eds.): Deep-Breath 2025, LNCS 16142, pp. 186–195, 2026.
https://doi.org/10.1007/978-3-032-05559-0_19

their ability to accurately reflect patient-specific anatomy and healing patterns, which can lead to disappointment after treatment.

Recent studies have utilised deep learning to predict post-operative outcomes based on pre-operative data. Freitas et al. [5] proposed a disentangled approach, while Montenegro et al. [12] introduced an inpainting framework that better preserves anatomical features. However, both methods primarily transfer asymmetries from other patients instead of generating personalised predictions, failing to capture traits such as scars.

This study introduces a conditional generative model based on Pix2Pix [7], extended with an encoder that integrates clinical metadata such as surgery type and cancer laterality, and explicitly preserves the skin colour and posture of the pre-operative image while manipulating the breasts. By conditioning the generator with pre-operative images and treatment-specific variables, the model offers realistic patient predictions of post-operative appearance.

2 Related Work

Generative models have enabled structured image translation across different domains. This section outlines important techniques for image-to-image translation and examines how these methods have been applied to predict post-operative outcomes, particularly in the context of breast cancer surgery.

2.1 Image-to-Image Translation Methods

Image-to-image translation (I2I) serves as a foundational technique for simulating surgical outcomes, enabling the transformation of pre-operative images into post-operative predictions. Most I2I methods are built on Generative Adversarial Networks (GANs) [6], which consist of a generator and a discriminator trained in opposition to produce increasingly realistic outputs. These methods can be categorised by their supervision (paired vs. unpaired), diversity (unimodal vs. multimodal), and domain scope (two-domain vs. multi-domain).

Among existing I2I models, Pix2Pix [7] uses conditional GANs with L1 loss to translate between aligned input-output pairs, making it well-suited for medical imaging tasks with paired data. CycleGAN [15] addresses unpaired settings using a cycle-consistency loss to enforce content preservation across domain mappings. To support diverse outputs, models such as BicycleGAN [16] and DRIT/DRIT++ [10] introduce latent codes and disentangled content-style representations, enabling multimodal generation. StarGAN [3] extends GANs to multi-domain tasks using label-based conditioning, while StarGAN v2 [4] further incorporates multimodal capabilities by leveraging style codes.

2.2 Post-operative Image Prediction

In post-operative outcome prediction for breast cancer treatment, Zolfagharnasab et al. [17] propose a regression-based approach to predict 3D breast surface deformations, but it lacks visual outputs and does not account for key

anatomical features. Freitas et al. [5] introduce a keypoint-based transformation method and a generative model with disentangled representations, which Montenegro et al. [11] refine using symmetry-based losses. In a parallel line of work, Montenegro et al. [12] propose an inpainting approach that reconstructs masked breast regions based on asymmetry cues. While more flexible, these generative methods still rely on template images from other patients, limiting their ability to generate fully personalised results. Furthermore, they focus solely on altering asymmetries, while ignoring other clinically-relevant traits like scars.

Chartier et al. [2] directly apply Pix2Pix to predict the outcomes of mammoplasty. However, they do not condition the model on surgery-related information, such as implant size, failing to capture surgery-specific traits. While it produces realistic results, it often alters unrelated features like skin tone and body pose. Despite these issues, it demonstrates the potential of supervised GANs for aesthetic outcome prediction.

3 Methodology

ACID-GAN is a conditional image-to-image translation model designed to generate personalised post-operative breast images (I_{post}) from pre-operative inputs (I_{pre}), while preserving identity traits such as skin tone, body shape, and posture. The architecture builds on the Pix2Pix framework [7], introducing clinical conditioning and semantic supervision through three main modules (Fig. 1):

- **Generator (G):** The generator predicts the post-operative image based on the pre-operative image and clinical information–specifically, surgery type and cancer laterality. The metadata is one-hot encoded, spatially replicated, and concatenated with the image as additional channels, which are then fed into a U-Net [13].
- **Discriminator (D):** The discriminator is responsible for promoting realism on the generated images. It follows a PatchGAN-style architecture and is trained on single images (pre-operative, post-operative and generated) to distinguish between real and fake appearances.
- **Encoder (E):** The encoder is a semantic classifier that receives image pairs: either (pre-, post-operative) images during the encoder training, or (pre-, generated) pairs during the generator training. Its task is to predict the surgery type and cancer laterality, enforcing alignment between the generated image and the conditioning vector. By decoupling this task from the discriminator, the model avoids entangling clinical semantics with domain-specific visual features such as lighting or skin tone.

Since the pre-operative and post-operative images are captured months or years apart, there are noticeable differences in illumination and skin tone between them. These variations can introduce biases during training, especially in how the model perceives colour and structure. To address these issues, the model is supervised using a combination of four complementary loss components, each designed to target a specific visual or anatomical aspect, as illustrated in Fig. 2.

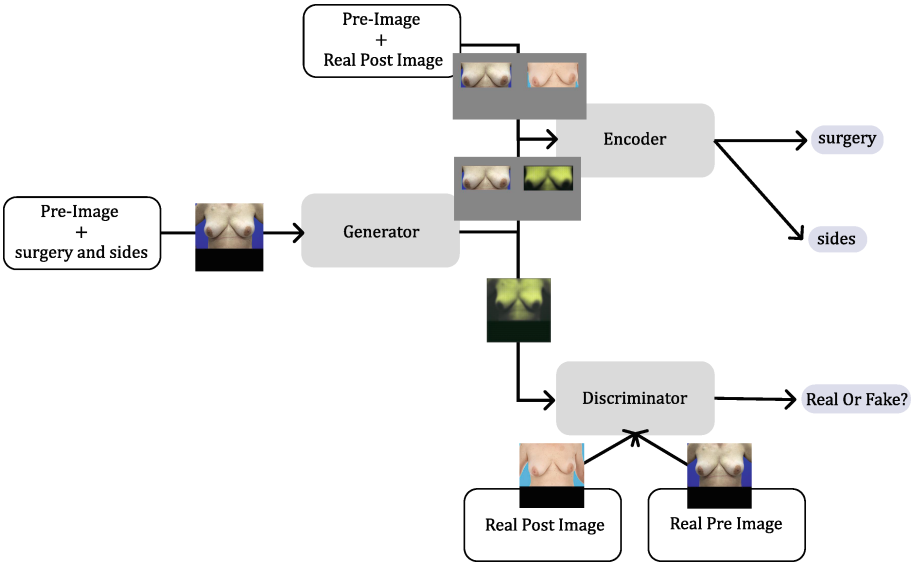


Fig. 1. Overview of the proposed ACID-GAN architecture with generator for image prediction, encoder for classification, and discriminator for realism

- **Colour Loss 1:** Compares average colour across patches using average pooling, promoting global colour harmony without affecting structure.

$$\mathcal{L}_{\text{colour1}} = \|\text{AvgPool}(G(I_{pre})) - \text{AvgPool}(I_{pre})\|_1 \tag{1}$$

- **Colour Loss 2:** Encourages preservation of skin tone and body shape by applying L1 loss between the generated and pre-operative images, excluding the breast region. The exclusion mask M_{in} is derived from a bounding box computed using anatomical keypoints of the breast contours.

$$\mathcal{L}_{\text{colour2}} = \|(1 - M_{\text{in}}) \odot (G(I_{pre}) - I_{pre})\|_1 \tag{2}$$

- **Structural Supervision:** Combines L1 and structural similarity image metric (SSIM) [14] on grayscale breast regions to enforce structural and perceptual similarity to the real post-operative image.

$$\begin{aligned}\mathcal{L}_{\text{structure}} = \; & \|\text{Gray}(M_{\text{in}} \odot G(I_{pre})) - \text{Gray}(M_{\text{in}} \odot I_{post})\|_1 \\ & + (1 - \text{SSIM}(\text{Gray}(M_{\text{in}} \odot G(I_{pre})), \text{Gray}(M_{\text{in}} \odot I_{post})))\end{aligned} \tag{3}$$

- **Segmentation Loss:** Applies cross-entropy (CE) between segmentation masks (Seg) of the generated and real images, obtained via a pretrained U-Net segmentation model. This ensures that anatomical features such as the breast and nipple are accurately aligned.

$$\mathcal{L}_{\text{seg}} = \text{CE}(\text{Seg}(G(I_{pre})), \text{Seg}(I_{post})) \tag{4}$$

Additionally, we use the typical adversarial loss of a GAN ($\mathcal{L}_{GAN}$) and binary cross-entropy between the predictions of the encoder and the type of surgery and breast side ($\mathcal{L}_{class}$). As such, the entire loss is depicted as Eq. 5, where λ represents parameters to control the relative importance of each term.

$$\begin{aligned}\mathcal{L}_G = \mathcal{L}_{\text{GAN}} + \lambda_{\text{L1}} \cdot \mathcal{L}_{\text{structure}} + \lambda_{\text{class}} \cdot \mathcal{L}_{\text{class}} \\ + \lambda_{\text{colour}} \cdot (\mathcal{L}_{\text{colour1}} + \mathcal{L}_{\text{colour2}}) + \lambda_{\text{seg}} \cdot \mathcal{L}_{\text{seg}}\end{aligned} \tag{5}$$

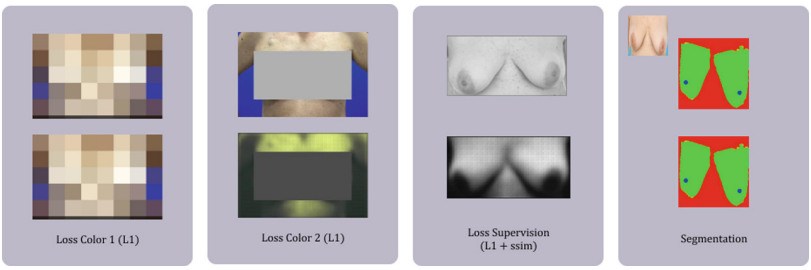


Fig. 2. Visual overview of the four loss components

The model was trained using the Adam optimiser (lr = 0.0002, $\beta_1 = 0.5$), with Xavier initialisation and a linear learning rate decay applied in the second half of training. To improve generalisation and prevent overfitting, dropout was applied in the generator, and data augmentations (e.g. horizontal flipping, mild affine transforms, and colour jitter) were used before providing the images to the discriminator. For SSIM supervision, a soft-edged mask with linear falloff replaced the rectangular binary mask, resulting in a mask that is one at the centre of the region and gradually tapers to 0 near the boundary, helping avoid sharp supervision discontinuities.

4 Experiments and Results

All experiments were conducted on a private dataset of pre- and post-operative frontal breast images collected at the Champalimaud Foundation as part of the CINDERELLA project[1]. A total of 1,038 image pairs were used, each consisting of a pre-operative and a post-operative image. These were split into 85% for training, 5% for validation, and 10% for testing.

To minimise variations in patient pose, zoom level, and background content, each image was cropped using a bounding box defined by anatomical keypoints outlining the breast. During cropping, a random zoom-out factor was applied for augmentation during training, along with a random horizontal flip. The dataset covers a range of surgical procedures, summarised in Table 1.

[1] https://cinderellaproject.eu/.

Table 1. Types of surgery and their frequencies in the dataset

Code	Description	Count
C1	Conservative Surgery	679
C2	Conservative Surgery with bilateral reduction	286
C3	Conservative Surgery with LD or LICAC/TDAP flaps	46
C4	Conservative Surgery with bilateral implants	27

To evaluate model performance, the following custom metrics were used, designed to capture both anatomical realism and visual consistency. These are referred to in all subsequent tables:

- **SSIM-Pre**: SSIM between the generated image and the original pre-operative input. Evaluates global consistency in body shape, pose, and lighting.
- **SSIM-Post**: SSIM with the ground truth post-operative image, computed in grayscale over the breast region. Measures fidelity to surgical changes like shape and scarring.
- **IoU-Breast**: Intersection over Union between predicted and ground truth breast segmentation masks.
- **IoU-Nipple**: Intersection over Union between predicted and ground truth nipple masks, capturing localised anatomical accuracy.
- **Colour1**: Patch-based Mean absolute error (MAE) comparing spatial averages between generated and pre-operative images. Penalises global colour shifts or desaturation.
- **Colour2**: MAE between generated and pre-operative images outside the breast region. Encourages consistency in skin tone and background.

Table 2 and Fig. 3 compare the results of the proposed method with the state-of-the-art Pix2Pix, BicycleGAN and StarGAN. As Pix2Pix and BicycleGAN are two-domain approaches that do not consider various surgical domains, we also include their results when conditioned on the surgery and cancer side. Furthermore, we measure the impact of removing the skip connections of the U-Net used as a generator and joining the encoder and discriminator into one network that simultaneously predicts the clinical conditions and image realism.

The Pix2Pix baseline model, when compared to its clinically conditioned variant, produced less consistent results. While it generated anatomically plausible transformations (Fig. 3), it often failed to reflect clinically meaningful distinctions. The conditioned variant, where one-hot encoded surgery and laterality labels were spatially replicated and concatenated with the input, yielded slightly improved predictions, particularly in the nipple IoU, suggesting better attention to surgical detail. Like Pix2Pix, BicycleGAN also benefited from clinical conditioning, though to a lesser extent. Concatenating metadata to the latent space led to slight improvements in $\text{SSIM}_{\text{post}}$ and nipple IoU, with BicycleGAN achieving the strongest segmentation performance overall. StarGAN, additionally supervised using SSIM, showed limited ability to reflect clinical transformations, with

Table 2. Quantitative results across all evaluated models

Model	SSIM (post) ↑	SSIM (pre) ↑	IOU (breast) ↑	IOU (nipple) ↑	Colour 1 ↓	Colour 2 ↓
Pix2Pix	0.75	0.73	0.77	0.05	0.07	0.07
BicycleGAN	0.76	0.74	0.78	0.07	0.07	0.07
Pix2Pix with conditions	0.76	0.75	0.78	**0.10**	0.07	0.07
BicycleGAN with conditions	0.77	0.77	0.79	0.08	0.07	0.06
StarGAN	0.76	0.77	0.75	0.05	0.06	0.06
Joint encoder/discriminator	**0.78**	0.84	0.79	0.08	0.03	0.01
ACID-GAN	**0.78**	**0.89**	**0.80**	**0.10**	**0.005**	**0.003**
ACID-GAN no skip connections	0.77	0.81	0.79	0.09	0.02	0.02

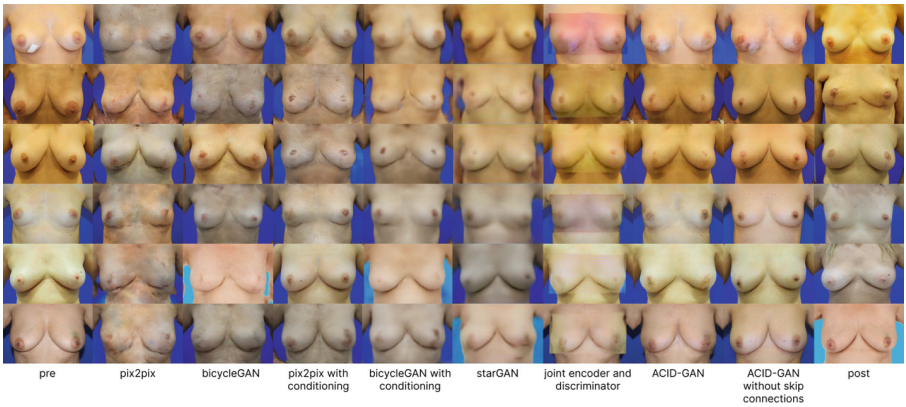


Fig. 3. Qualitative comparison across methods. Each row shows a different patient, with the pre-operative image, predictions from the models, and the ground-truth post-operative image. The proposed ACID-GAN produces the most realistic and clinically consistent outputs, preserving skin tone and body shape while accurately reflecting surgical changes. Baselines often struggle with colour consistency and anatomical detail

nipple IoU remaining unchanged at 0.05. Overall, these three models exhibited difficulties in preserving consistent skin tone, as reflected in their Colour 1 and Colour 2 scores and in the visual results.

The best performance was achieved by the proposed ACID-GAN model. With skip connections, it reached the highest SSIM_{pre} and good colour consistency, highlighting strong preservation of patient identity. Removing skip connections improved anatomical detail but reduced colour fidelity, revealing a trade-off between visual sharpness and appearance consistency (visible in Fig. 3). Merging the encoder and discriminator into a single module yielded images with visible boxes around the breast area and lower scores than the separated encoder design, since SSIM_{pre} dropped from 0.89 to 0.84, IOU from 0.1 to 0.08, and Colour 1 increased from 0.005 to 0.03, likely due to conflicting optimisation pressures.

Finally, the results of the proposed model with and without skip connections on each surgery type, highlighted in Table 3 and Fig. 4, show limitations in the less frequent surgery types (C3 and C4), whose outputs often resembled those of

Table 3. Per-surgery breakdown of results for the two ACID-GAN variants

	Surgery, Count	SSIM (post) ↑	SSIM (pre) ↑	IOU (breast) ↑	IOU (nipple) ↓	Colour 1 ↓	Colour 2 ↓
ACID-GAN	C1, 75	**0.79**	**0.90**	**0.79**	**0.12**	**0.005**	**0.003**
	C2, 24	**0.74**	**0.87**	**0.80**	0.03	**0.006**	**0.002**
	C3, 4	0.79	**0.87**	0.79	0.01	**0.006**	**0.003**
	C4, 4	**0.82**	**0.89**	**0.78**	**0.13**	**0.005**	**0.003**
ACID-GAN without skip connections	C1, 75	0.78	0.81	**0.79**	0.09	0.02	0.02
	C2, 24	**0.74**	0.80	0.79	**0.04**	0.01	0.02
	C3, 4	**0.81**	0.78	**0.80**	**0.16**	0.02	0.02
	C4, 4	**0.82**	0.80	0.77	0.08	0.02	0.02

more common classes like C2. This may be due to two factors: the low number of C3/C4 samples and anatomical bias in the dataset, where such surgeries are typically associated with smaller breast volumes. Increasing the classification loss weight for underrepresented procedures could improve their distinctiveness, but risks destabilising training or overfitting due to data scarcity.

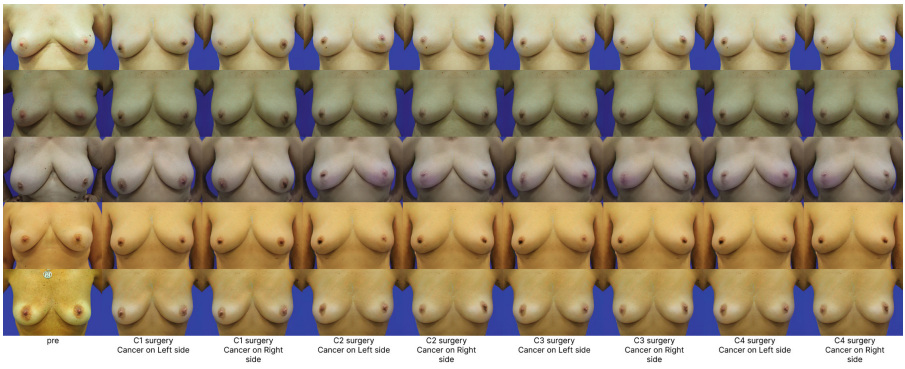


Fig. 4. Predictions from ACID-GAN without skip connections across different surgery types (C1–C4) and cancer sides, using the same pre-operative image. Changes in symmetry, volume, and nipple position are noticeable, reflecting surgery-specific outcomes

5 Conclusion

This study presents a clinically informed image-to-image translation framework for simulating aesthetic outcomes in breast cancer locoregional treatment. By extending Pix2Pix with semantic conditioning and anatomically-guided loss functions, the model is able to generate personalised post-operative predictions from pre-operative images. Results on a real-world clinical dataset confirm that conditioning on surgery metadata improves the alignment of outputs with clinical context, and that the choice of architectural components impacts the trade-off between visual fidelity and structural clarity. Despite promising results, challenges remain, particularly in modelling less frequent surgical types. Moreover,

the current conditioning is limited to surgery type and laterality, which may restrict generalisability across more diverse patient profiles. Future work will explore the incorporation of richer patient metadata (e.g., age, cup size) using more flexible conditioning mechanisms, such as language-based encoders, and improved handling of underrepresented surgery types. Validation through clinical expert feedback and integration into decision-support tools will also be essential to assess the perceptual and practical utility of the model in real-world settings.

Acknowledgement. This work has received funding from the European Union's Horizon Europe research and innovation programme under the Grant Agreement 101057389-CINDERELLA project, and from the FCT - Foundation for Science and Technology Portugal, within PhD grant 2022.14516.BD.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Cardoso, J., Silva, W., Cardoso, M.: Evolution, current challenges, and future possibilities in the objective assessment of aesthetic outcome of breast cancer locoregional treatment. Breast **49**, 123–130 (2020). https://doi.org/10.1016/j.breast.2019.11.006
2. Chartier, C., Watt, A., Lin, O., Chandawarkar, A., Lee, J., Hall-Findlay, E.: Breastgan: artificial intelligence-enabled breast augmentation simulation. In: Aesthetic Surgery Journal Open Forum, vol. 4, p. ojab052. Oxford University Press US (2022)
3. Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S., Choo, J.: Stargan: unified generative adversarial networks for multi-domain image-to-image translation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8789–8797 (2018)
4. Choi, Y., Uh, Y., Yoo, J., Ha, J.W.: Stargan v2: diverse image synthesis for multiple domains. In: Proceedings of the IEEECVF Conference on Computer Vision and Pattern Recognition, pp. 8188–8197 (2020)
5. Freitas, N., Montenegro, H., Cardoso, M.J., Cardoso, J.S.: Reproducing asymmetries caused by breast cancer treatment in pre-operative breast images. In: 2024 IEEE International Symposium on Biomedical Imaging (ISBI), pp. 1–5. IEEE (2024)
6. Goodfellow, I., et al.: Generative adversarial networks. Commun. ACM **63**(11), 139–144 (2020)
7. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1125–1134 (2017)
8. Kaidar-Person, O., et al.: Evaluating the ability of an artificial-intelligence cloud-based platform designed to provide information prior to locoregional therapy for breast cancer in improving patient's satisfaction with therapy: the cinderella trial. PLoS ONE **18**(8), e0289365 (2023)
9. Kim, M.K., et al.: Effect of cosmetic outcome on quality of life after breast cancer surgery. Eur. J. Surg. Oncol. **41**(3), 426–432 (2015)

10. Lee, H.Y., Tseng, H.Y., Huang, J.B., Singh, M., Yang, M.H.: Diverse image-to-image translation via disentangled representations. In: Proceedings of the European Conference on Computer vision (ECCV), pp. 35–51 (2018)
11. Montenegro, H., Cardoso, M.J., Cardoso, J.S.: A disentangled approach to predict the aesthetic outcomes of breast cancer treatment. In: European Conference on Computer Vision, pp. 311–327. Springer (2025)
12. Montenegro, H., Cardoso, M.J., Cardoso, J.S.: An inpainting approach to manipulate asymmetry in pre-operative breast images. arXiv preprint arXiv:2502.05652 (2025)
13. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
14. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004)
15. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2223–2232 (2017)
16. Zhu, J.Y., et al.: Toward multimodal image-to-image translation. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
17. Zolfagharnasab, H., et al.: A regression model for predicting shape deformation after breast conserving surgery. Sensors **18**(1), 167 (2018)



BreastSegNet: Multi-label Segmentation of Breast MRI

Qihang Li[1], Jichen Yang[1], Yaqian Chen[2(✉)], Yuwen Chen[2], Hanxue Gu[2], Lars J. Grimm[3], and Maciej A. Mazurowski[1,2,3,4]

[1] Department of Biostatistics and Bioinformatics, Duke University, Durham, USA
[2] Department of Electrical and Computer Engineering, Duke University, Durham, USA
yc591@duke.edu
[3] Department of Radiology, Duke University Medical Center, Durham, USA
[4] Department of Computer Science, Duke University, Durham, USA

Abstract. Breast MRI provides high-resolution imaging critical for breast cancer screening and preoperative staging. However, existing segmentation methods for breast MRI remain limited in scope, often focusing on only a few anatomical structures, such as fibroglandular tissue or tumors, and do not cover the full range of tissues seen in scans. This narrows their utility for quantitative analysis. In this study, we present **BreastSegNet**, a multi-label segmentation algorithm for breast MRI that covers nine anatomical labels: fibroglandular tissue (FGT), vessel, muscle, bone, lesion, lymph node, heart, liver, and implant. We manually annotated a large set of 1123 MRI slices capturing these structures with detailed review and correction from an expert radiologist. Additionally, we benchmark nine segmentation models, including U-Net, SwinUNet, UNet++, SAM, MedSAM, and nnU-Net with multiple ResNet-based encoders. Among them, nnU-Net ResEncM achieves the highest average Dice scores of **0.694** across all labels. It performs especially well on heart, liver, muscle, FGT, and bone, with Dice scores exceeding **0.73**, and approaching **0.90** for heart and liver. All model code and weights are publicly available, and we plan to release the data at a later date.

Keywords: Breast Segmentation · Breast Composition · Body Composition

1 Introduction

Breast magnetic resonance imaging (MRI) is one of the primary imaging modalities for breast cancer staging and high-risk screening, providing superior contrast resolution and high sensitivity for early disease identification [15]. Multiple studies have recently demonstrated its superiority in breast cancer risk prediction, early detection, and diagnostic evaluation [6,19], compared to traditional imaging modalities such as mammography.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
T. Zhang et al. (Eds.): Deep-Breath 2025, LNCS 16142, pp. 196–205, 2026.
https://doi.org/10.1007/978-3-032-05559-0_20

Despite these benefits, quantitative research on breast MRI remains limited, largely due to the lack of publicly available segmentation methods [9]. A high-accuracy segmentation model that provides rich semantic information can facilitate a variety of quantitative research, including the analysis based on breast composition [2] and registration [5].

General-purpose medical image segmentation models, such as TotalSegmentator [1], overlook labels specific to breast MRI, like fibroglandular tissue (FGT). A previously proposed breast MRI segmentation model proposed by Lew et al. [3,13] provides only FGT, vessel, and breast labels. The lack of necessary labels precludes the direct extraction of important body composition parameters—such as muscle quality and bone density—from breast MRI [16,18], thereby reducing the clinical applicability of these models.

To mitigate this gap, this study aims to provide a publicly available segmentation model and dataset that offers nine labels, including FGT, vessel, muscle, bone, lesion, lymph node, heart, liver, and implant. The main contributions to this work are:

1. We provide an automated, multi-class breast MRI segmentation model with nine anatomical labels, evaluated by the Dice similarity coefficient.
2. We developed and evaluated nine segmentation architectures, with the top-performing model, attaining an average Dice score greater than **0.694** for all nine labels.
3. We release the model code and pretrained weights on https://github.com/mazurowski-lab/BreastSegNet. We also describe the a breast MRI dataset and corresponding annotations which we plan to release at a later date.

2 Dataset

2.1 Data Selection

The initial cohort was collected from female patients who underwent bilateral breast MRI with and without contrast, along with paired radiology reports within our health system between 2014 and 2023. In total, we identified 5,865 patients with 10,836 breast MRI exams. We then downloaded and de-identified these exams, and those that failed to download or were missing in the system were excluded. After this process, a total of 10,083 exams from 5,793 patients were available for further use.

From the initial cohort, we randomly selected 108 exams from different patients and divided them into two groups of 58 and 50 for training and testing, respectively. These volumes were identified manually. The data distribution for the training and testing sets is illustrated in Fig. 1.

2.2 Annotation Pipeline

The whole annotation process was carried out by four researchers without formal radiology training and one expert fellowship-trained breast radiologist who

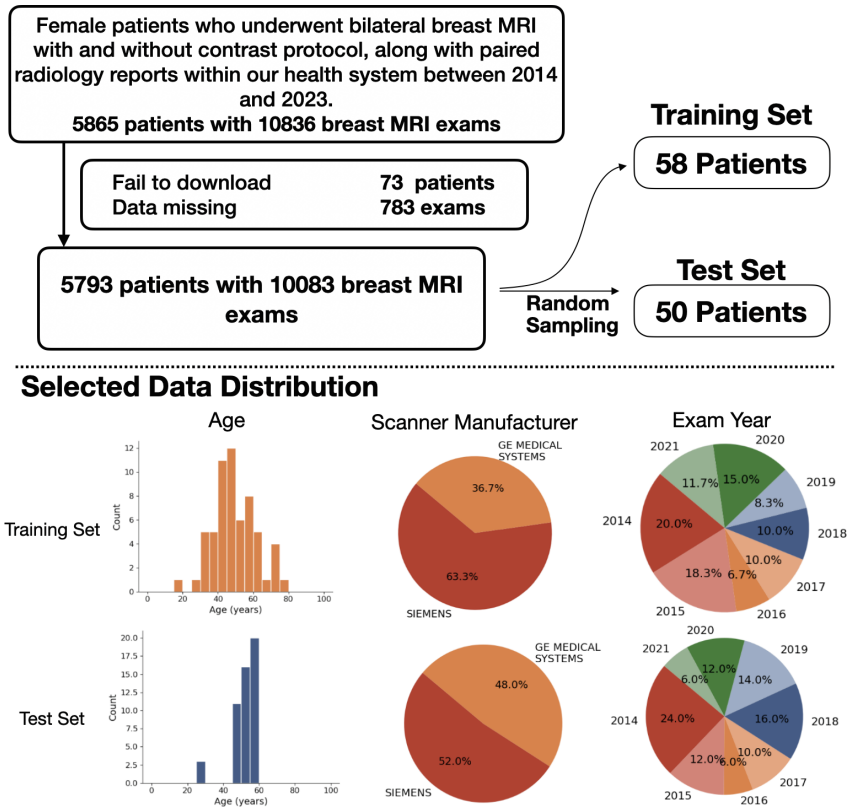


Fig. 1. The source data population, and the selected data distribution.

supervised the process. As illustrated in Fig. 2, we employed an iterative, model-assisted annotation workflow for breast MRI image annotation, with radiologist oversight at each stage. The annotation contains nine labels: tissue, vessel, muscle, bone, lesion, lymph node, heart, liver, and implant. To maximize the distinction between different labels, we selected the first post-contrast MRI sequence for annotation. The process consisted of multiple rounds of model development and manual refinement:

Initially, four patient MRIs were manually annotated by trained annotators, followed by radiologist review and approval to ensure accuracy. The annotations were applied every 10 slices. The annotators assigned each pixel to one of the nine labels, or background, to the best of their knowledge. The radiologist reviewed each annotated slice and provided feedback if they found any annotations inaccurate. Each slice was iteratively revised until the radiologist approved the annotation. These annotated cases were used to train a preliminary segmentation model (Temp Model A), which provided initial predictions for the next set of images.

Subsequently, predictions from Temp Model A were generated for 12 additional patient MRIs. These predictions were manually revised by annotators and reviewed by a radiologist to produce refined annotations. The updated dataset was then used to train a second model (Temp Model B).

In the final iteration, Temp Model B was applied to 58 new patient MRIs. The resulting segmentations were again manually corrected by researchers and approved by the radiologist. These curated annotations were used to train the final model.

For independent evaluation, a separate test set of 50 patient MRIs was manually annotated without model assistance. For each volume, 3 random slices were randomly selected from the breast region and annotated. All annotations in this test set underwent radiologist review to serve as the ground truth for model performance assessment.

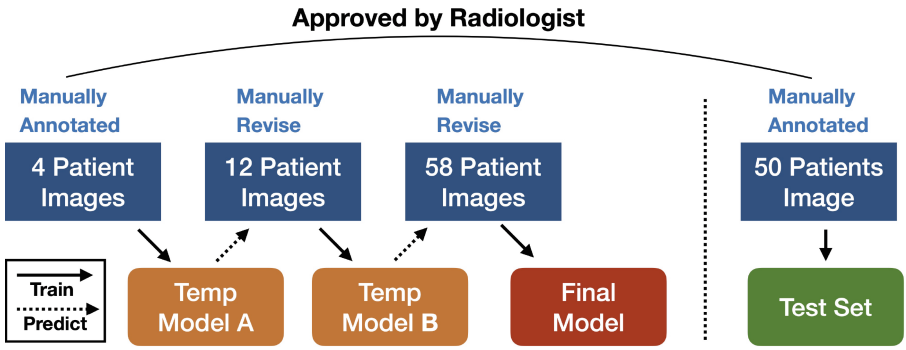


Fig. 2. The data annotation pipeline.

2.3 Segmentation Details

This section provides detailed definitions of the nine labels supported by our model, including FGT, vessel, muscle, bone, lesion, lymph node, heart, liver, and implant.

The FGT, vessel, muscle, bone, and lymph node, annotations strictly adhered to anatomical definitions, as these classifications present no inherent ambiguity. Lesion refers to all abnormal findings within the breast that are not artificially implanted, encompassing masses, non-mass enhancements, and metastatic/ cancerous lymph nodes. Due to the presence of artifacts, the boundaries of the heart and liver are often indistinct on MRI, but we have made every effort to differentiate them. In this annotation, implant specifically denotes materials used in breast augmentation procedures, including saline or silicone implants. Other artificial objects, such as implantable infusion ports, were not labeled.

Figure 3 presents example slices with overlaid annotations (*orange*) for the nine segmented labels. For each label, two randomly selected slices are shown.

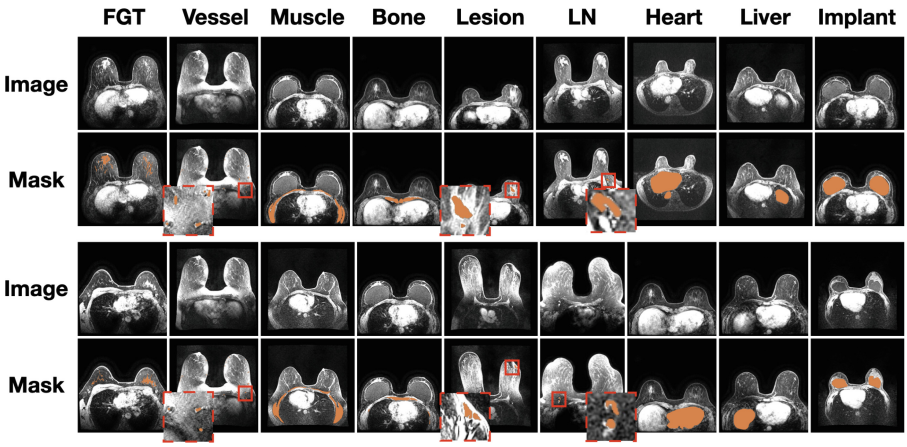


Fig. 3. Annotation visualization for nine anatomical labels: For each label, two representative slices are shown: the original image (rows 1 and 3) and its corresponding segmentation overlaid in *orange* (rows 2 and 4). For small or sparse structures (Vessel, Lesion, LN), zoomed-in views are provided in dashed red boxes, with the original regions marked by solid red boxes. (Color figure online)

The first two rows correspond to the original image and the image with overlaid segmentation mask from the first selected slice, while the third and fourth rows show the same for a second slice.

3 Methods

3.1 Segmentation Methods

To select the best-performing algorithm for our task, we explored nine segmentation methods, which include U-Net, SwinUNet, U-Net++, fine-tuned SAM, fine-tuned MedSAM, nnU-Net, nnU-Net ResEncM, nnU-Net ResEncL, and nnU-Net ResEncXL. These methods are grouped into three categories for better comparison: (1) the U-Net-based category, which includes U-Net, SwinUNet, and U-Net++; (2) the foundation model category, comprising fine-tuned SAM and fine-tuned MedSAM; and (3) the nnU-Net-based category, consisting of nnU-Net and its variants with ResEncM, ResEncL, and ResEncXL encoder [4,10–12,14,17,20]. The detailed descriptions of the models shown follow.

U-Net. U-Net is one of the most commonly used benchmarks for medical image segmentation tasks. Compared to traditional convolutional networks, U-Net features a symmetric encoderdecoder structure that helps preserve spatial resolution during upsampling. SwinUNet builds on the original U-Net architecture by integrating the Swin Transformer into the encoder, enabling the capture of long-range dependencies. U-Net++, on the other hand, introduces nested and dense skip connections, which help reduce the semantic gap between encoder and decoder features.

Foundation Model. Foundation model fine-tuning has become a recent trend in medical segmentation tasks due to its strong adaptability and generalizability. It has demonstrated state-of-the-art performance across multiple medical segmentation tasks [8]. Among these models, the Segment Anything Model (SAM) and MedSAM are two of the most commonly used foundation models for medical image segmentation. SAM is trained on a large corpus of natural images and demonstrates strong zero-shot segmentation capabilities. MedSAM is further adapted for medical applications by being pretrained on a diverse set of medical imaging data, making it widely applicable across segmentation tasks.

NnU-Net-Based. We separate the nnU-Net algorithms from the U-Net category to highlight the performance differences introduced by different nnU-Net planners. The original nnU-Net extends the U-Net architecture by incorporating automated configuration of preprocessing, architecture adaptation, training, and postprocessing pipelines, demonstrating significant performance improvements over the vanilla U-Net. In its recent update, nnU-Net further introduces enhancements to the U-Net baseline, emphasizing the importance of using advanced CNN architectures such as ResNet. The ResEncM, ResEncL, and ResEncXL planners represent nnU-Net variants with progressively deeper ResNet encoder backbones—ResNet-34, ResNet-50, and ResNet-101.

3.2 Evaluation Metrics

We employed the Dice coefficient as the metric to evaluate the model. Dice is one of the most commonly utilized metrics, measuring the image overlap between predicted and ground truth regions, and reflecting the spatial agreement of segmentation boundaries.

The formula for the Dice coefficient is Eq. (1), where A represents the model-predicted mask and the B represents the set of pixels in the ground truth. Dice ranges from 0 to 1, where a value of 1 indicates perfect overlap and 0 signifies no overlap [7].

$$\text{Dice} = \frac{2|A \cap B|}{|A| + |B|} \tag{1}$$

4 Experiments and Results

4.1 Implementation Details

This study utilized 58 volumetric datasets from 58 patients for training. Of these, 812 slices from 48 volumes were allocated to the training set and 161 slices from 10 volumes to the validation set.

All models were configured for 2D prediction. For the nnU-Net series, default slice-wise normalization was applied. To ensure consistency, slice-wise normalization was also used for the other models. The fine-tuned models, including SAM and MedSAM, were trained for 100 epochs on our datasets. All other models

were trained for a total of 1000 epochs, and the model achieving the best performance on the validation set was selected as the final model. All models converged during training.

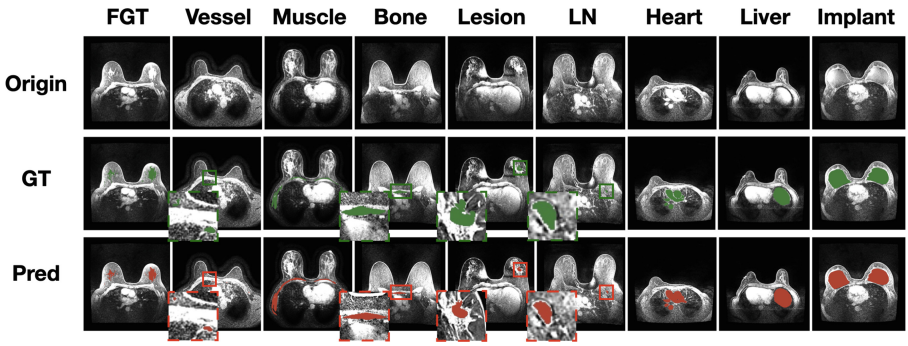


Fig. 4. Qualitative comparison: Segmentation results of nnU-Net ResEncM, the best-performing model, compared to ground truth (GT) across nine anatomical labels. Rows show the original image, GT (*green*), and prediction (*red*). Zoomed-in views are shown for small structures (Vessel, Lesion, LN). (Color figure online)

4.2 Results

To evaluate segmentation performance, we computed the average 2D Dice score across all slices containing non-zero labels for each anatomical label across the nine selected models. The quantitative results are detailed in Table 1, and a qualitative comparison between the prediction of nnU-Net ResEncM, which demonstrates the highest average performance, and the ground truth is shown in Fig. 4.

As demonstrated in Table 1, nnU-Net and its ResEnc variants significantly outperform the models in the UNet and Foundation Model categories for this task, with the best-performing models belonging to nnU-Net and its ResEnc variants across all nine labels. Among the four models in the nnU-Net-based category, nnU-Net ResEncM exhibits the highest average performance, achieving an average Dice score of **0.694**.

Segmentation performance varies notably across anatomical labels, with some structures consistently achieving higher Dice scores. For example, all models in the nnU-Net-based category achieve Dice scores above **0.73** for the FGT and bone labels, over **0.8** for the muscle label, and approach **0.9** for the heart and liver labels. For some more challenging regions such as vessel, lesion, and implant, the performance varies significantly across the models, with the best scores reaching **0.58**, **0.66**, and **0.81**, respectively.

Notably, lymph nodes exhibit the lowest segmentation performance among the nine labels. We discuss potential reasons for this limitation and plans to address it in the discussion section.

Table 1. Segmentation results for nine models, evaluated by Dice (↑). The highest score for each label among all models is highlighted in bold. The right-most column reports the mean Dice for each label across every model. Acronyms: FGT - fibroglandular tissue.

Models in UNet and Foundation Model Categories					
Composition	UNet	SwinUNet	UNet++	SAM	MedSAM
FGT	0.66	0.69	0.69	0.60	0.56
Vessel	0.41	0.46	0.50	0.23	0.18
Muscle	0.68	0.69	0.74	0.70	0.64
Bone	0.61	0.62	0.67	0.51	0.42
Lesion	0.00	0.01	0.07	0.05	0.01
Lymph node	0.00	0.00	0.03	0.14	0.05
Heart	0.85	0.85	0.83	0.86	0.84
Liver	0.42	0.25	0.26	0.71	0.54
Implant	0.60	0.18	0.53	0.80	0.74
Overall	0.470	0.417	0.480	0.511	0.442
Models in nnU-Net-based Category					
Composition	nnU-Net	ResEncM	ResEncL	ResEncXL	Avg
FGT	0.74	**0.75**	0.74	0.74	0.69
Vessel	**0.58**	0.57	0.56	0.57	0.45
Muscle	0.80	**0.81**	0.80	0.81	0.74
Bone	**0.73**	**0.73**	**0.73**	**0.73**	0.64
Lesion	0.38	**0.66**	0.38	0.63	0.24
LN	0.13	0.12	**0.16**	0.10	0.08
Heart	**0.92**	0.91	0.91	0.91	0.88
Liver	0.89	**0.90**	**0.90**	**0.90**	0.64
Implant	**0.81**	0.80	0.76	0.69	0.66
Overall	0.664	**0.694**	0.660	0.675	0.56

For qualitative evaluation, Fig. 4 presents the comparison between ground truth (GT) and the predictions (Pred) from nnU-Net ResEncM across all nine anatomical labels. The original MRI slice is shown in the first row, followed by the GT mask (*green*) and the predicted mask (*red*). For large and clearly defined structures such as FGT, muscle, bone, heart, and liver, the model shows strong agreement with the ground truth, with minimal boundary discrepancy. For smaller and more challenging labels like vessel, lesion, and lymph node, visual differences become more apparent. For these small or sparse regions, zoomed-in boxes are provided to better assess model performance.

5 Discussion

In this study, we proposed **BreastSegNet**, a multi-label segmentation network for breast MRIs, to address the limitations of prior models, which typically target limited structures. By incorporating nine anatomical labels, including clinically important regions such as fibroglandular tissue, vessel, muscle, bone, lesion, lymph node, heart, liver, and implant, our model enables more comprehensive downstream analysis in breast imaging. To ensure that our data results in the most accurate model, evaluate nine segmentation frameworks, with nnU-Net ResEncM demonstrating the highest Dice score of 0.694 averaged across all nine anatomical labels. Notably, segmentation accuracy is especially strong for large anatomical structures such as the heart, liver, muscle, FGT, and bone, each achieving Dice scores above 0.73 and approaching 0.90 for the heart and liver.

The limitation of our model is the lower segmentation performance for smaller or less visually distinct structures, especially lymph nodes. We identify three key factors contributing to this lower performance: (1) Lymph nodes are rare. Only 15 out of 150 test slices have lymph nodes. This strong data imbalance makes the model prone not to predict them, as the original loss treats matched absence in prediction and ground truth as zero loss. (2) Lymph nodes have similar attentuation values as blood vessels due to their vascular supply, making it easy for the model to confuse them. (3) Lymph nodes, especially non-cancerous lymph nodes, are usually small structures, so slight mislabeling in prediction leads to a dramatic Dice drop.

References

1. Akinci D'Antonoli, T., et al.: Totalsegmentator mri: robust sequence-independent segmentation of multiple anatomic structures in mri. Radiology **314**(2), e241613 (2025)
2. Boyd, N.F., et al.: Breast tissue composition and susceptibility to breast cancer. J. Natl Cancer Inst. **102**(16), 1224–1237 (2010)
3. Buda, M., et al.: Data from the breast cancer screening–digital breast tomosynthesis (bcs-dbt). Data from The Cancer Imaging Archive (2020)
4. Cao, H. et al.: Swin-Unet: Unet-like pure transformer for medical image segmentation. In: Karlinsky, L., Michaeli, T., Nishino, K. (eds.) ECCV 2022. LNCS, vol. 13803. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-25066-8_9
5. Chen, Y., et al.: Guidedmorph: two-stage deformable registration for breast mri. arXiv preprint arXiv:2505.13414 (2025)
6. Chen, Y., et al.: Breast density in mri: an ai-based quantification and relationship to assessment in mammography. arXiv preprint arXiv:2504.15192 (2025)
7. Dice, L.R.: Measures of the amount of ecologic association between species. Ecology **26**(3), 297–302 (1945)
8. Gu, H., et al.: Segmentanybone: a universal model that segments any bone at any location on mri. Med. Image Anal., 103469 (2025)
9. Gubern-Mérida, A., Kallenberg, M., Mann, R.M., Marti, R., Karssemeijer, N.: Breast segmentation and density estimation in breast mri: a fully automatic framework. IEEE J. Biomed. Health Inform. **19**(1), 349–357 (2014)

10. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: Nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nat. Methods **18**(2), 203–211 (2021)
11. Isensee, F., et al.: nnu-net revisited: a call for rigorous validation in 3d medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 488–498. Springer (2024). https://doi.org/10.1007/978-3-031-72114-4_47
12. Kirillov, A., et al.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4015–4026 (2023)
13. Lew, C.O., et al.: A publicly available deep learning model and dataset for segmentation of breast, fibroglandular tissue, and vessels in breast mri. Sci. Rep. **14**(1), 5383 (2024)
14. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nat. Commun. **15**(1), 654 (2024)
15. Mann, R.M., Kuhl, C.K., Kinkel, K., Boetes, C.: Breast mri: guidelines from the european society of breast imaging. Eur. Radiol. **18**, 1307–1318 (2008)
16. McGregor, R.A., Cameron-Smith, D., Poppitt, S.D.: It is not just muscle mass: a review of muscle quality, composition and metabolism during ageing as determinants of muscle function and mobility in later life. Longevity Healthspan **3**, 1–8 (2014)
17. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
18. Shepherd, J.A., Ng, B.K., Sommer, M.J., Heymsfield, S.B.: Body composition by dxa. Bone **104**, 101–105 (2017)
19. Wekking, D., Porcu, M., Silva, P., Saba, L., Scartozzi, M., Solinas, C.: Breast mri: clinical indications, recommendations, and future applications in breast cancer diagnosis. Curr. Oncol. Rep. **25**(4), 257–267 (2023)
20. Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J.: Unet++: a nested u-net architecture for medical image segmentation. In: Stoyanov, D., et al. (eds.) DLMIA/ML-CDS -2018. LNCS, vol. 11045, pp. 3–11. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00889-5_1



Anatomy-Preserving Counterfactual Edits in Breast MRI via Guided Diffusion

Bulat Maksudov[1(✉)], Kathleen M. Curran[2], and Alessandra Mileo[1]

[1] School of Computing, Dublin City University, Dublin, Ireland
bulat.maksudov2@mail.dcu.ie
[2] School of Medicine, University College Dublin, Dublin, Ireland

Abstract. Dynamic contrast-enhanced breast MRI is highly sensitive but difficult to interpret. We ask whether counterfactual edits produced by a latent diffusion model can reveal image cues linked to tumor presence while preserving anatomy. Starting from a latent diffusion backbone, we fine-tune the U-Net denoiser on the MAMA-MIA dataset and generate slice-level edits via DDIM inversion and null-text inversion. We study two operations: curtailment (suppression of tumor evidence) and exaggeration (amplification or addition of tumor-like features). Semantic impact is quantified with a frozen 3D nnU-Net trained on expert tumor masks. Across cohorts, curtailment yields a consistent 18–43% mean reduction in predicted tumor extent, whereas exaggeration is less reliable, highlighting an asymmetry between subtractive and additive edits. These results suggest diffusion-based counterfactuals provide interpretable, anatomy-preserving "what-if" views that complement saliency maps for model auditing, hypothesis generation, and training (Code and trained models are available at https://github.com/Luab/breast-mri-counterfactuals.)

Keywords: Explainable AI · Diffusion models · Counterfactual explanations

1 Introduction

Analyzing breast MRI images remains challenging even for experienced radiologists [10]. Decision support tools powered by AI can assist in triage, preliminary segmentation, and diagnostic interpretation, ultimately reducing radiologists' workload and improving diagnostic accuracy and patient outcomes.

Explainable AI (xAI) refers to a set of techniques designed to make model predictions more interpretable. Common xAI methods include saliency maps, attention mechanisms, and counterfactual explanations. These methods typically produce visual overlays in the same pixel space as the input images, aiming to highlight regions that influenced the model's prediction. However, current approaches often lack specificity and suffer from poor spatial localization, limiting their clinical utility [1,14].

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
T. Zhang et al. (Eds.): Deep-Breath 2025, LNCS 16142, pp. 206–215, 2026.
https://doi.org/10.1007/978-3-032-05559-0_21

Diffusion models have recently emerged as powerful generative tools capable of producing high-fidelity and diverse outputs. They offer fine control over generation through conditioning and guidance mechanisms [4,9,13]. These properties have enabled a range of image editing techniques that can subtly alter image content while preserving realism.

We build on this foundation to generate visual counterfactual explanations for breast MRI - that is, modified versions of input images that illustrate hypothetical changes relevant to clinical decision-making. Specifically, we define two types of counterfactual transformations: *curtailment*, which reduces or suppresses a particular entity (e.g., remove or suppress tumor evidence in the image), and *exaggeration*, which amplifies or adds tumor-like features. These operations aim to answer "what if" questions such as: *What if this slice contained healthy tissue?* or *What if this tumor were more conspicuous on contrast-enhanced MRI?*, i.e. suggestive of higher malignancy, while preserving the anatomical and contextual integrity of the original image.

In this work, we propose a pipeline for generating and evaluating counterfactual explanations in breast MRI using a fine-tuned diffusion model. Our key contributions are:

- We introduce an entity-preserving counterfactual generation mechanism that enables precise, fine-grained manipulation of target tumor features while maintaining overall anatomical consistency.
- We propose a segmentation-based evaluation protocol complemented by *Fréchet Radiomics Distance (FRD)*, to quantitatively assess the clinical fidelity and realism of generated counterfactuals, grounded in clinically meaningful segmentation tasks.

1.1 Background - Diffusion Model Adaptation

Originally diffusion models were introduced to operate in pixel space of images [3], where z_0 corresponds to the image itself. This approach turned out to be computationally expensive and a new class of Latent Diffusion models was introduced [12]. Rather than applying diffusion process to the image, it maps the original image to the latent space of the Autoencoder model and maps back to the image space at the end of the process.

For domain specific application, the common approach is to fine-tune pretrained model. For example in medical imaging in [15], authors fine-tune only textual embedding by introducing new tokens, while in [7] the textual conditioning part stays frozen, while the denosing model is fine-tuned. In both cases the embeddings of the autoencoder model remain unchanged. We follow the same approach, which is training only the denoising model, while keeping the rest of the model weights frozen. The overall training pipeline is presented in Fig. 1.

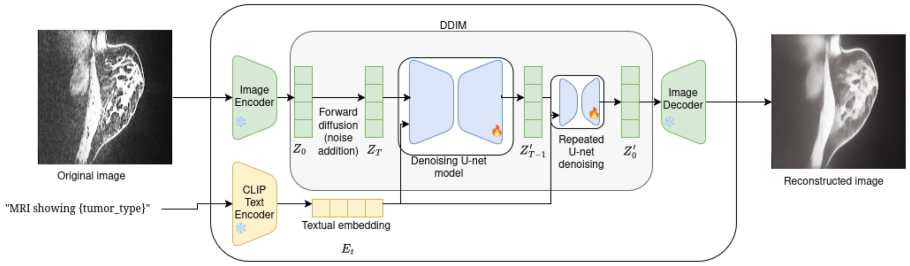


Fig. 1. Training pipeline for counterfactual image generation using a diffusion model. The original MRI is encoded into a latent vector and diffused with added noise. A CLIP text encoder provides conditioning. A fine-tuned U-Net performs DDIM-based denoising, after which the latent is decoded to reconstruct or modify the image.

2 Methodology

2.1 Generating Counterfactual Explanations

Guided diffusion models work by mapping a random noise vector z_t and conditioning vector E_t into image vector z_0. This denosing is performed sequentially by a network ϵ_θ, which is trained to predict true noise residual ϵ. [3]

Conditioning vector is obtained by embedding prompt via CLIP textual model [11] and z_t is obtained by adding noise to the original image vector z_0 according timestamp t. At inference time a noisy vector z_T is gradually denoised by predicting noise residual for T timestamps by denosing network. To accurately reconstruct a given noise sample, a Denoising Diffusion Implicit Model (DDIM) [13] sampling is used:

$$z_{t-1} = \sqrt{\frac{\alpha_{t-1}}{\alpha_t}} z_t + \left(\sqrt{\frac{1}{\alpha_{t-1}} - 1} - \sqrt{\frac{1}{\alpha_t} - 1} \right) \cdot \varepsilon_\theta(z_t, t, E_t). \tag{1}$$

DDIM Inversion. Equation (1) gives the *reverse* DDIM update ($z_t \to z_{t-1}$). For counterfactual editing we must first *invert* a real scan x_0 into the noise regime. Re-arranging Eq. (1) yields the deterministic *forward* step

$$z_{t+1} = \sqrt{\frac{\alpha_{t+1}}{\alpha_t}}\, z_t + \left(\sqrt{\frac{1}{\alpha_{t+1}} - 1} - \sqrt{\frac{1}{\alpha_t} - 1} \right) \varepsilon_\theta\big(z_t, t, E_t\big), \qquad t = 0, \ldots, T-1, \tag{2}$$

where $z_0 = E_{\text{image}}(x_0)$ and ε_θ is the fine-tuned denoiser. Iterating Eq. (2) drives the latent to a noisy point z_T that remains *exactly* reconstructible under Eq. (1).

Classifier-Free Guidance (CFG). To steer sampling without an external classifier we adopt classifier-free guidance [4]. At each reverse step t the denoiser predicts two residuals:

$$\varepsilon_\theta\big(z_t, t, E_t^c\big) \text{ (conditional)}, \quad \varepsilon_\theta\big(z_t, t, E_t^u\big) \text{ (unconditional; } E_t^u = E_t(\text{""})\text{)}.$$

They are blended via the guidance scale w:

$$\hat{\varepsilon}_\theta(z_t, t) = w\,\varepsilon_\theta(z_t, t, E_t^c) + (1 - w)\,\varepsilon_\theta(z_t, t, E_t^u). \tag{3}$$

Replacing ε_θ by $\hat{\varepsilon}_\theta$ in Eq. (1) yields the guided reverse update that realises the desired counterfactual transformation.

Guided Reverse Sampling. Starting from z_T, we apply classifier-free guidance to obtain the counterfactual image. The guided residual $\hat{\varepsilon}_\theta(z_t, t, E_t)$ is inserted back into Eq. (1). Subsequent null-text optimisation further tightens the reconstruction fidelity; see Eq. (4).

Null-Text Inversion. Large guidance scales ($w \gg 0$) may distort subject-specific anatomy, compromising identity preservation. To mitigate this, we adopt the null-text inversion scheme of [9]. A separate unconditional embedding $\varnothing_t^*$ is optimised at every reverse step while all network weights remain frozen, see Fig. 2. For each timestep t we solve

$$\varnothing_t^* = \arg\min_{\varnothing_t} \left\| z_{t-1} - f_{\text{DDIM}}\big(z_t, t, \hat{\varepsilon}_\theta(z_t, t, E_t, \varnothing_t)\big) \right\|_2^2, \tag{4}$$

with

$$\hat{\varepsilon}_\theta(z_t, t, E_t, \varnothing_t) = w\,\varepsilon_\theta(z_t, t, E_t) + (1 - w)\,\varepsilon_\theta(z_t, t, \varnothing_t).$$

Using the optimised sequence $\{\varnothing_t^*\}_{t=1}^T$, we replace the unconditional embedding E_t^u in Eq. (3) at every timestep. This injects the per-step null-text trajectory directly into the guided reverse pass, ensuring identity-preserving counterfactuals while realising the intended semantic edits (see Fig. 3).

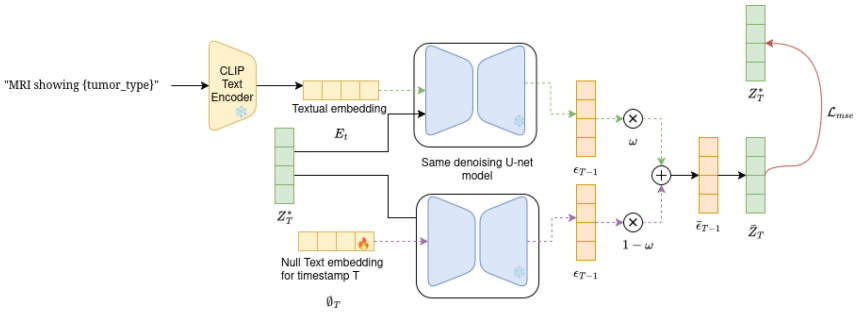


Fig. 2. Null-text inversion pipeline. A noisy latent z_T^* is denoised twice using a shared U-Net model: once with the text embedding E_t, and once with the null-text embedding $\varnothing_T$. The weighted combination of the two predictions, controlled by guidance scale ω, yields the denoised latent $\tilde{z}_{T-1}$, which is compared to the ground truth z_{T-1}^* via MSE loss. 🔥 denotes trainable parameters; ❄ indicates frozen weights.

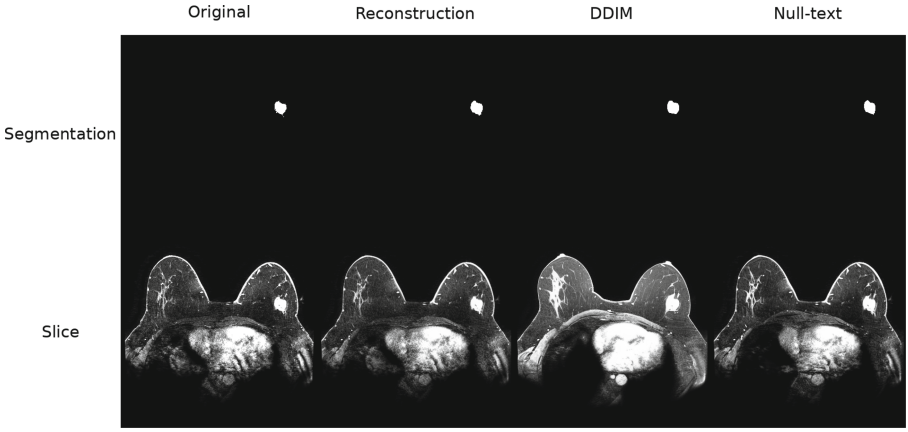


Fig. 3. Example outputs for DUKE patient 507, slice 42 with luminal tumor subtype. Observation: Compared to null-text inversion, DDIM inversion introduces broader changes beyond the lesion, affecting surrounding anatomical structures - which may reduce anatomical fidelity in clinical interpretation.

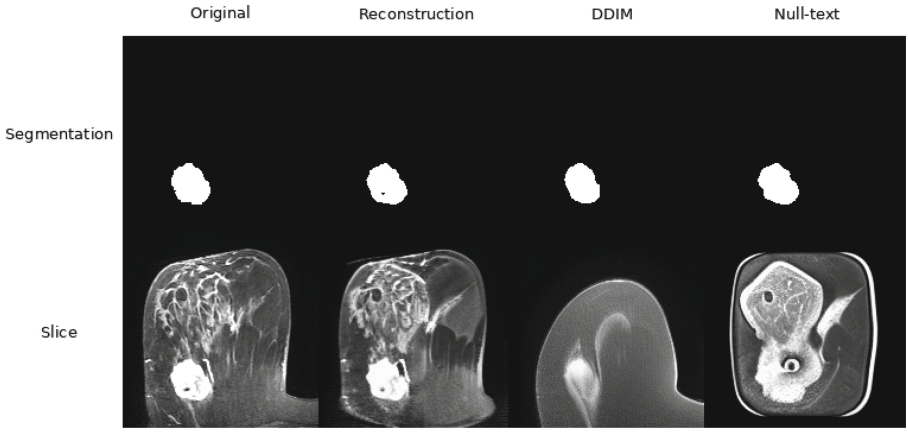


Fig. 4. Example outputs for ISPY2 patient 613188, slice 55 with triple negative tumor subtype. Observation: This case demonstrates model instability under domain shift, where even the basic autoencoding reconstruction substantially alters anatomical characteristics.

3 Experimental Setup

3.1 Dataset

All experiments use the MAMA–MIA benchmark [2], a multi-centre breast DCE-MRI set of 1 506 T_1-weighted studies drawn from the I-SPY1, I-SPY2, DUKE and NACT cohorts. Each study is paired with a 3-D tumor segmentation refined by 16 breast-imaging specialists and comes with 49 harmonised clinical variables

(age, subtype, pCR, survival, etc.). Baseline nnU-Net weights released by the dataset authors which we utilize for counterfactual evaluation.

3.2 Preprocessing

All volumes were resampled to isotropic $1\,\mathrm{mm}^3$ spacing. Global Z-score normalization was applied across all phases, followed by intensity clipping to the 1st–99th percentile and linear mapping to $[-1, 1]$. Expert segmentations were resampled to the same grid using nearest-neighbor interpolation. Slices with poor signal quality—e.g., zero intensity, low variance, low dynamic range, or >95% background—were excluded.

3.3 Segmentation to Caption

Each slice that passes the preprocessing stage (Sect. 3.2) is paired with a short natural-language prompt that later serves as conditioning text for the diffusion model. The prompt is derived from the expert segmentation mask and corresponding metadata. For every pre-processed slice we form a short prompt: if its expert mask contains no tumor voxels we write "MRI showing normal tissue"; otherwise we write "MRI showing <tumor_type>′′, where tumor_type is the per-patient subtype recorded in the metadata (triple-negative, luminal-a/b, luminal, or HER2-enriched/pure)

3.4 Training Specific Information

Hardware and Initialisation. All experiments ran on a single NVIDIA RTX 4090 GPU (24 GB). The network was initialised from the public `Stable-Diffusion-v1-5` [12] checkpoint. Both the VAE decoder and the CLIP text encoder were kept frozen throughout the study.

Editing and Null-Text Optimisation. For DDIM inversion we propagated the latent through 50 forward timesteps. At each timestep the null-text embedding was refined by 20 Adam steps with learning rate 1×10^{-2}; the rate decayed linearly as $\mathrm{lr}' = \mathrm{lr}\,(1 - i/N_t)$ where i is the timestep index. Gradients were clipped at norm 1.0, and optimisation stopped early if the per-step residual fell below 1×10^{-5}. The final counterfactual was sampled with 50 DDIM reverse steps and classifier-free guidance scale 7.5, identical to the scale used during the null-text loss computation.

4 Evaluation

4.1 Segmentation Based Evaluation

Quantitatively assessing counterfactual images in breast MRI is difficult, because the image that would appear after a hypothetical edit is never observed. Without such a reference we look for a stand-in that is likely to move when clinically relevant anatomy changes while staying mostly stable otherwise. tumor

segmentation suits these requirements: (i) the output mask has a clear geometric meaning, and (ii) large public datasets provide expert annotations on which to train a reliable model. Therefore, if the predicted tumor extent decreases after a "curtailment" edit, or increases after an "exaggeration" edit, we interpret this shift as indirect evidence that the counterfactual has had the intended semantic impact. Prior work in explainable AI reports similar use of downstream tasks as surrogate metrics [8,16] however, to our knowledge this is the first use within diffusion-based MRI editing.

Volume Insertion Protocol. All counterfactual manipulations are applied to individual slices after the common preprocessing pipeline 3.2. To feed a slice back into the 3-D network we re-insert the transformed slice at its original position within the preprocessed $H \times W \times D$ volume, leaving all other voxels unchanged. The resulting 3-D stack is processed through nnUNet [5].

Evaluation Conditions. Each patient slice is analysed under four conditions: (i) Original (no change), (ii) Reconstruction obtained by DDIM inversion with guidance weight $\omega = 0$, (iii) plain DDIM inversion followed by a classifier-free-guidance reverse pass, and (iv) our full null-text counterfactual that adds null-text optimisation on top of classifier-free guidance. For each condition we compute the mean number of tumor pixels per slice and report it in absolute terms as well as relative to the original. A lower (higher) relative number after curtailment (exaggeration) is taken as a successful steer.

Table 1. Counterfactual editing results. Mean segmentation size in pixels and mean *relative* size (ratio to Original, in parentheses). For **curtailment** we seek *lower* values (↓); null-text inversion yields the largest average reduction across all four cohorts. For **exaggeration** the goal is *higher* values (↑); gains over the Reconstruction baseline are marginal, indicating room for improvement. Best value per evaluation is in **bold**.

	Baseline		Curtailment ↓		Exaggeration ↑	
Dataset	Orig.	Recon.	DDIM	Null	DDIM	Null
DUKE	575	517 (0.90)	474 (0.83)	**317 (0.55)**	537 (0.93)	**557 (0.97)**
ISPY1	925	921 (0.99)	857 (0.89)	**653 (0.71)**	948 (1.02)	**972 (1.05)**
ISPY2	488	436 (0.89)	299 (0.61)	**291 (0.60)**	361 (0.74)	408 (0.84)
NACT	966	960 (0.99)	782 (0.81)	**884 (0.91)**	957 (0.99)	960 (0.99)

Table 2. Fréchet Radiomics Distance (FRD)between original slices and edited variants (lower is better). Bold indicates the lowest value per column.

Condition vs. Original	Curtailment FRD ↓	Exaggeration FRD ↓
Reconstruction	**8.27**	8.36
DDIM inversion	10.95	10.74
Null-text inversion	9.38	**8.20**

5 Results and Discussion

Quantitative Performance. The quantitative results for curtailment and exaggeration experiments are presented in Table 1 respectively. For curtailment on every cohort the nnU-net baseline shows a double-digit percentage drop in predicted tumor pixels after null-text editing. The strongest mean reduction is observed on DUKE, the smallest on NACT, but all lie within an 18–43% band, confirming that the edits remove cues the segmentation network relies on.

For exaggeration we observe a far less uniform picture: some cohorts exhibit a modest pixel-count increase, others even a decrease. The overall trend suggests that additive edits are intrinsically harder than subtractive ones.

FRD Analysis. We employ Fréchet Radiomics Distance [6], which was introduced to measure quality of generative medical imaging models by using radiomics features. The results are provided in Table 2. On curtailment, DDIM increases FRD by +32.3% over reconstruction (10.95 vs. 8.27), whereas null-text limits the shift to +13.5% (9.38). Surprisingly, for exaggeration, null-text attains the lowest FRD (8.20), 1.9% below reconstruction (8.36) and 23.7% lower than DDIM (10.74). These results indicate that null-text inversion better preserves the radiomics distribution and by extension clinically meaningful properties of the originals.

Failure Modes. Out-of-distribution slices–e.g. heavily cropped single-breast views-sometimes collapse even in the reconstruction step, indicating that the latent prior struggles when the acquisition deviates too far from training data. See Fig. 4.

5.1 Future Work

We envision future work in improving the conditioning mechanisms by utilizing Control-Net-style adapter layers, combined with anatomical atlases would enable fine-grained control of the generative process. We would also like to acknowledge the difficulty of evaluating counterfactual medical images in terms of visual plausibility. Current image-quality metrics (FID, SSIM) correlate poorly with clinical realism. We envision a reader study with panel of radiologists to asses anatomical and clinical plausibility of induced changes.

6 Conclusion

We have demonstrated that a lightweight fine-tuned latent diffusion model can generate slice-level counterfactual edits readable by clinicians for breast DCE-MRI while preserving the surrounding anatomy. When the editing is subtractive (curtailment), feeding the modified slice back into a frozen 3D nnU-Net consistently lowers the predicted tumor area by 18- 43% in four cohorts, providing indirect yet quantitative evidence that diagnostically relevant cues are removed.

In contrast, tumor exaggeration yields only marginal or inconsistent gains, indicating room for improvement in additive editing. Although occasional anatomy drift and domain-shift collapses remain, the proposed pipeline already offers a practical tool for probing "what-if" scenarios in high-stakes imaging and lays the groundwork for future extensions such as volumetric editing, stronger conditioning, and perceptual realism metrics tailored to MRI.

Acknowledgements. This publication has been the result of research conducted with the financial support of Science Foundation Ireland under Grant number 18/CRT/6183. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

Disclosure of Interests.. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Arun, N., et al.: Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging. Radiology: Artifi. Intell. **3**(6), e200267 (2021). https://doi.org/10.1148/ryai.2021200267
2. Garrucho, L., et al.: A large-scale multicenter breast cancer dce-mri benchmark dataset with expert segmentations. Sci. Data **12**(1), 453 (2025). https://doi.org/10.1038/s41597-025-04707-4
3. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Adv. Neural Inform. Process. Syst. **33**, 6840–6851. (2020)
4. Ho, J., Salimans, T.: Classifier-free diffusion guidance (2022). https://arxiv.org/abs/2207.12598
5. Isensee, F., Jaeger, P., Kohl, S., Petersen, J., Maier-Hein, K.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nat. Methods **18**, 1–9 (2021). https://doi.org/10.1038/s41592-020-01008-z
6. Konz, N., et al.: Fréchet radiomic distance (frd): a versatile metric for comparing medical imaging datasets (2025). https://arxiv.org/abs/2412.01496
7. Kumar, A., Kriz, A., Havaei, M., Arbel, T.: PRISM: high-resolution & precise counterfactual medical image generation using language-guided stable diffusion. In: Medical Imaging with Deep Learning (2025). https://openreview.net/forum?id=UpJMAlZNuo
8. Mehta, R., et al.: Cf-seg: Counterfactuals meet segmentation (2025). https://arxiv.org/abs/2506.16213
9. Mokady, R., Hertz, A., Aberman, K., Pritch, Y., Cohen-Or, D.: Null-text inversion for editing real images using guided diffusion models. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6038–6047. IEEE Computer Society, Los Alamitos, CA, USA (2023). https://doi.org/10.1109/CVPR52729.2023.00585
10. Pötsch, N., et al.: Breast mri: does a clinical decision algorithm outweigh reader experience? European Radiol. **32** (2022). https://doi.org/10.1007/s00330-022-09015-8

11. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning (2021). https://api.semanticscholar.org/CorpusID:231591445
12. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-Resolution Image Synthesis with Latent Diffusion Models . In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10674–10685. IEEE Computer Society, Los Alamitos, CA, USA (2022). https://doi.org/10.1109/CVPR52688.2022.01042, https://doi.ieeecomputersociety.org/10.1109/CVPR52688.2022.01042
13. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: International Conference on Learning Representations (2021). https://openreview.net/forum?id=St1giarCHLP
14. Viviano, J.D., Simpson, B., Dutil, F., Bengio, Y., Cohen, J.P.: Saliency is a possible red herring when diagnosing poor generalization. In: International Conference on Learning Representations (2021). https://openreview.net/forum?id=c9-WeM-ceB
15. de Wilde, B., Saha, A., de Rooij, M., Huisman, H., Litjens, G.: Medical diffusion on a budget: Textual inversion for medical image generation. In: Medical Imaging with Deep Learning (2024). https://openreview.net/forum?id=J0zEnfU3Ow
16. Xia, T., Ribeiro, F.D.S., Rasal, R.R., Kori, A., Mehta, R., Glocker, B.: Decoupled classifier-free guidance for counterfactual diffusion models (2025). https://arxiv.org/abs/2506.14399



Uncertainty-Aware Cross-Modal Attention for Breast Cancer Classification in Ultrasound Imaging

Aiman Farooq[1(✉)], Chandisha Das[2], and Deepak Mishra[1]

[1] Indian Institute of Technology Jodhpur, Jodhpur 342030, India
farooq.1@iitj.ac.in
[2] Future Institute of Technology, Sonarpur 700154, India

Abstract. We propose an uncertainty-aware cross-modal attention framework for breast cancer classification using ultrasound imaging and structured clinical data. Our model incorporates a Clinical Knowledge Gate (CKG) that adaptively weights clinical features based on medical domain relevance, and an uncertainty-aware attention mechanism that dynamically fuses modalities based on confidence estimates. We employ contrastive learning to enhance alignment between modalities and address class imbalance through an adaptive focal loss with uncertainty regularization. Evaluated on the Breast-Lesions-USG dataset, our method outperforms traditional and deep learning baselines, achieving 90% accuracy and 98% recall. Grad-CAM visualizations confirm alignment with radiological markers, while uncertainty estimates improve interpretability and clinical trust.

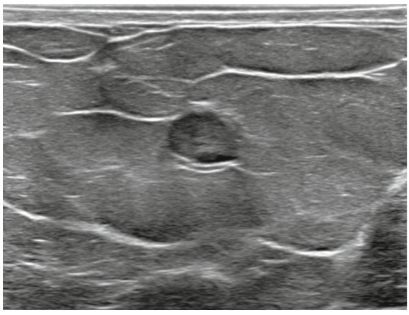
Keywords: Breast cancer classification · Uncertainty quantification · Cross-modal attention · Clinical knowledge integration · Medical imaging

1 Introduction

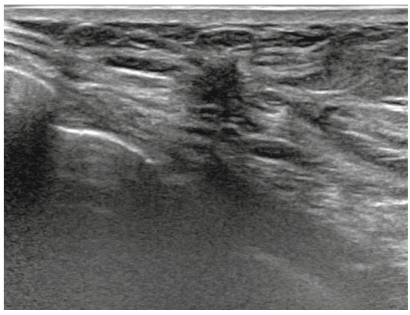
Breast cancer is the second most common cancer globally, and early detection plays a pivotal role in improving patient outcomes [26]. Screening programs have significantly enhanced survival rates, with five-year survival exceeding 90% when diagnosed at localized stages, compared to only 28% for metastatic disease [3]. Among diagnostic tools, ultrasound imaging is widely used due to its non-invasive nature, accessibility, and real-time imaging capabilities [12]. Despite its utility, accurately distinguishing between benign and malignant breast lesions on ultrasound remains a complex task—even for experienced radiologists—due to substantial overlap in visual features and inherent subjectivity in interpretation [16]. As shown in Fig. 1, malignant lesions may exhibit irregular shapes, spiculated margins, and heterogeneous echogenicity, while benign lesions typically have smooth contours and homogeneous internal structure. However, these traits often overlap, contributing to diagnostic uncertainty [20].

A. Farooq and C. Das—Contributed equally.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
T. Zhang et al. (Eds.): Deep-Breath 2025, LNCS 16142, pp. 216–225, 2026.
https://doi.org/10.1007/978-3-032-05559-0_22



(a) Benign lesion with smooth margins



(b) Malignant lesion with irregular shape

Fig. 1. Ultrasound images showing different types of breast lesions: (a) Benign lesion, (b) Malignant lesion

Recent advances in deep learning have shown promise for automating breast cancer classification [1,15,24]. Early approaches using conventional convolutional neural networks (CNNs) [10] have evolved into more sophisticated models incorporating attention mechanisms and multi-scale features [25]. Comprehensive surveys of radiomics and AI applications in breast imaging [31]have highlighted the potential for automated diagnostic systems while identifying key challenges in clinical implementation For instance, Raza et al. [24] proposed DeepBreastCancerNet utilizing ensemble learning, Islam et al. [15] introduced a multi-scale feature extraction framework, and Alzoubi et al. [1] developed advanced architectures for robust classification.

Despite these advances, most models rely solely on imaging data, neglecting critical clinical information routinely used by radiologists. Clinical features such as patient demographics (e.g., age, family history), lesion characteristics (e.g., size, location, morphology), and radiological assessments (e.g., BI-RADS categories [4], vascular patterns) provide complementary diagnostic value [5]. Multimodal approaches integrating imaging and clinical data have shown potential to enhance diagnostic performance [5]. For example, Yala et al. [30] demonstrated that combining mammographic findings with clinical risk factors improved breast cancer risk prediction. In contrast, Wang et al. [29] reported superior performance of multi-modal fusion strategies across various medical imaging tasks. However, simplistic strategies—such as direct feature concatenation or basic weighted averaging—without accounting for modality reliability or prediction uncertainty [18]. The growing emphasis on uncertainty quantification in medical AI aims to increase trust and clinical acceptance [6].

In summary, current approaches face several limitations: they often exclude valuable clinical features, lack uncertainty quantification essential for clinical decision-making, and provide limited interpretability [18]. Clinical features such as patient age, BI-RADS scores [4], and lesion descriptors should be systematically integrated to improve diagnostic robustness and trustworthiness. To

address these gaps, we propose an uncertainty-aware cross-modal attention framework for breast cancer classification that integrates ultrasound imaging with structured clinical data while providing interpretable confidence estimates. Our contributions are threefold:

- We introduce a Clinical Knowledge Gate (CKG) that adaptively weights clinical features based on medical domain relevance, enabling more meaningful fusion than naive concatenation.
- We propose an uncertainty-aware cross-modal attention mechanism that estimates confidence for each modality and adjusts fusion weights based on reliability.
- We incorporate a comprehensive training pipeline with contrastive learning for cross-modal alignment and adaptive focal loss to address class imbalance, improving accuracy and uncertainty calibration.

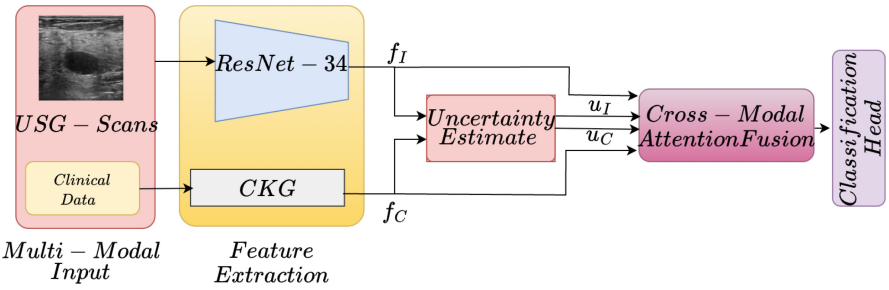


Fig. 2. Overview of the proposed uncertainty-aware cross-modal attention framework. The model processes ultrasound images through a ResNet-34 backbone and clinical features through a Clinical Knowledge Gate (CKG). An uncertainty-aware attention mechanism estimates modality-specific uncertainties and dynamically fuses the representations

2 Methodology

Figure 2 illustrates the overall architecture of our proposed framework. Given an ultrasound image $I \in \mathbb{R}^{H \times W}$ and a clinical feature vector $C \in \mathbb{R}^d$, the objective is to predict a binary label $y \in \{0, 1\}$ corresponding to benign or malignant diagnosis, while also estimating the associated predictive uncertainty. Formally, the framework learns a mapping $f(I, C) \rightarrow (y, u)$, where u denotes the uncertainty associated with the prediction. Unlike conventional models that yield only deterministic outputs, our framework explicitly quantifies prediction confidence—a critical requirement for safe clinical deployment [6].

2.1 Clinical Knowledge Gate

Traditional fusion approaches treat all clinical features equally, typically employing simple concatenation or weighted averaging strategies [29]. However, medical domain knowledge indicates that certain features (e.g., BI-RADS score, lesion shape, margin characteristics) are more diagnostically significant than others in determining malignancy risk [20]. Radiological studies have consistently shown that irregular margins, spiculated borders, and heterogeneous echogenicity strongly predict malignancy [27].

We introduce a Clinical Knowledge Gate (CKG) incorporating this medical expertise through learnable domain-informed weighting. The CKG mechanism initializes feature importance weights based on established clinical guidelines and allows adaptive refinement during training:

$$\mathbf{w}_{medical} = [\alpha_{BIRADS}, \alpha_{shape}, \alpha_{margin}, \alpha_{echo}, \alpha_{age}, \alpha_{composition}] \tag{1}$$

where α values are learnable parameters initialized with medical priors derived from radiological literature [7]. The adaptive gating mechanism computes feature-specific attention weights by considering both the input clinical features and learned medical knowledge:

$$\mathbf{g} = \sigma(W_g \mathbf{C} + \mathbf{b}_g) \tag{2}$$

$$\mathbf{c}_{gated} = \mathbf{g} \odot \mathbf{w}_{medical} \odot \mathbf{c} \tag{3}$$

where σ is the sigmoid activation function, $W_g \in \mathbb{R}^{d \times d}$ are learned transformation weights, $\mathbf{b}_g \in \mathbb{R}^d$ represents bias terms, and $\odot$ denotes element-wise multiplication. This mechanism allows the model to dynamically adjust feature importance based on the specific case context while maintaining consistency with established medical knowledge.

2.2 Uncertainty-Aware Cross-Modal Attention

Cross-modal attention mechanisms have demonstrated success in various multimodal learning tasks [19], but their application to medical imaging with uncertainty quantification remains underexplored. We develop a novel attention mechanism that quantifies uncertainty for each modality and uses these estimates to dynamically weight the fusion process, addressing the challenge that different modalities may exhibit varying degrees of reliability across different cases.

For image features $\mathbf{f}_I \in \mathbb{R}^{d_I}$ extracted from a pre-trained ResNet-34 backbone and clinical features $\mathbf{f}_C \in \mathbb{R}^{d_C}$ processed through the clinical knowledge gate, we compute modality-specific uncertainties using dedicated uncertainty estimation networks:

$$u_I = \sigma(MLP_I(\mathbf{f}_I)), \quad u_C = \sigma(MLP_C(\mathbf{f}_C)) \tag{4}$$

where MLP_I and MLP_C are multi-layer perceptrons with hidden dimensions [256, 128, 1] that learn to predict the epistemic uncertainty associated with each modality, with uncertainty values bounded between 0 and 1.

The dynamic fusion weights are computed by considering both feature content and their associated uncertainties through a joint attention mechanism:

$$\mathbf{w} = \text{softmax}(MLP_{weight}([\mathbf{f}_I; \mathbf{f}_C; u_I; u_C])) \tag{5}$$

where $[\cdot;\cdot]$ denotes concatenation and MLP_{weight} is a fusion network that learns optimal weighting strategies. The final fused representation incorporates confidence scaling to down-weight unreliable modalities:

$$\mathbf{f}_{fused} = w_I \cdot (1 - u_I) \cdot \mathbf{f}_I + w_C \cdot (1 - u_C) \cdot \mathbf{f}_C \tag{6}$$

This formulation ensures that modalities with high uncertainty (approaching 1) contribute less to the final representation, while confident predictions (uncertainty near 0) receive full weighting.

2.3 Contrastive Cross-Modal Alignment

Effective multi-modal learning requires establishing semantic correspondence between different modalities [23]. To improve the alignment between image and clinical embeddings, we employ contrastive learning that encourages matched image-clinical pairs to have similar representations while pushing apart unmatched pairs.

We project both modalities to a shared embedding space of dimension 128 using dedicated projection heads ϕ_I and ϕ_C, followed by L2 normalization to obtain unit vectors:

$$\mathbf{z}_I = \frac{\phi_I(\mathbf{f}_I)}{||\phi_I(\mathbf{f}_I)||_2}, \quad \mathbf{z}_C = \frac{\phi_C(\mathbf{f}_C)}{||\phi_C(\mathbf{f}_C)||_2} \tag{7}$$

The contrastive loss employs the InfoNCE framework [21], which has proven effective for learning discriminative representations:

$$\mathcal{L}_{contrastive} = -\log \frac{\exp(\text{sim}(\mathbf{z}_I, \mathbf{z}_C)/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(\mathbf{z}_I, \mathbf{z}_C^{(j)})/\tau)} \tag{8}$$

where $\text{sim}(\cdot,\cdot)$ represents cosine similarity, $\tau = 0.07$ is the temperature parameter that controls the concentration of the distribution, N is the batch size, and j indexes all clinical embeddings in the batch.

2.4 Adaptive Focal Loss with Uncertainty Regularization

Medical datasets often exhibit significant class imbalance, with benign cases typically outnumbering malignant ones [2]. Additionally, incorporating uncertainty estimates requires appropriate regularization to ensure meaningful confidence predictions. We design a composite loss function that addresses both challenges:

$$\mathcal{L}_{total} = \mathcal{L}_{focal} + \lambda_u \mathcal{L}_{uncertainty} + \lambda_c \mathcal{L}_{contrastive} \tag{9}$$

where $\lambda_u = 0.1$ and $\lambda_c = 0.1$ are hyperparameters that balance the different loss components.

The adaptive focal loss [17] addresses class imbalance by down-weighting easy examples and focusing learning on hard cases:

$$\mathcal{L}_{focal} = -\alpha_t (1 - p_t)^\gamma \log(p_t) \tag{10}$$

where p_t is the predicted probability for the true class, α_t is a class-specific weighting factor computed from inverse class frequencies, and $\gamma = 2$ is the focusing parameter. The uncertainty regularization term encourages reasonable uncertainty levels while preventing overconfident or overly uncertain predictions:

$$\mathcal{L}_{uncertainty} = |u_I - 0.3| + |u_C - 0.3| + |u_{pred} - 0.2| \tag{11}$$

where u_{pred} represents the overall prediction uncertainty computed as $u_{pred} = \sqrt{w_I^2 u_I^2 + w_C^2 u_C^2}$. The target uncertainty values (0.3 for modalities, 0.2 for predictions) are set based on empirical analysis of optimal uncertainty ranges that balance informativeness with calibration.

3 Experimental Setup

3.1 Dataset

We evaluate our approach on the Breast-Lesions-USG dataset from The Cancer Imaging Archive [22], a curated benchmark for ultrasound-based breast lesion analysis. The dataset contains 256 breast ultrasound scans, with both benign (64.1%) and malignant (35.9%) lesions. Each case includes a grayscale ultrasound image (224×224 pixels) and 14 clinical features, including patient demographics, BI-RADS category, lesion characteristics (shape, margin, echogenicity), tissue composition, and radiological findings.

3.2 Implementation Details

Our framework uses ResNet-34 as the image backbone, with image embeddings projected to 128 dimensions and clinical features to 64 dimensions. We train for 100 epochs using the Adam optimizer with a learning rate of 1e-4 and a weight decay of 1e-5. The uncertainty and contrastive loss weights are set to $\lambda_u = 0.1$ and $\lambda_c = 0.1$. We employ 5-fold stratified cross-validation for robust evaluation, where each fold uses approximately 80% of data for training and 20% for validation while preserving the original class distribution across all splits. We assess performance using accuracy, precision, recall, F1-score, and AUC. For uncertainty evaluation, we analyze calibration, correlation between uncertainty and prediction errors, and coverage at different confidence thresholds.

We conduct ablation studies to analyze the contribution of each component Table 2).

Table 1. Classification performance comparison on breast ultrasound dataset across different fusion and modalities.

Method	Modality	Accuracy	Precision	Recall
KNN [9]	Clinical	0.85 ± 0.05	0.80 ± 0.10	0.74 ± 0.14
Logistic Regression [13]	Clinical	0.86 ± 0.05	0.80 ± 0.10	0.85 ± 0.13
Random Forest [8]	Clinical	0.87 ± 0.05	0.82 ± 0.09	0.86 ± 0.08
ResNet50 [11]	Image	0.73 ± 0.06	0.49 ± 0.09	0.36 ± 0.08
ResNet34 [11]	Image	0.73 ± 0.05	0.48 ± 0.10	0.33 ± 0.07
EfficientNet [28]	Image	0.81 ± 0.04	0.76 ± 0.07	0.73 ± 0.06
Early Fusion [14]	Image + Clinical	0.86 ± 0.05	0.92 ± 0.06	0.67 ± 0.14
Late Fusion [14]	Image + Clinical	0.80 ± 0.09	0.91 ± 0.07	0.51 ± 0.26
Ours	Image + Clinical	**0.90 ± 0.07**	**0.93 ± 0.08**	**0.98 ± 0.07**

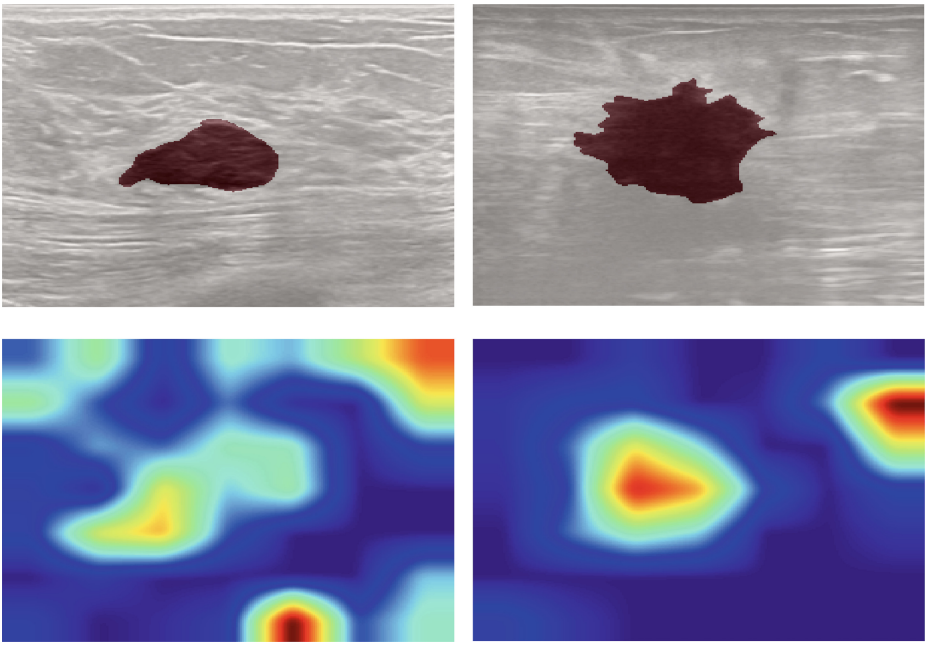


Fig. 3. The first column illustrates a benign lesion, while the second column presents a malignant lesion. The top row shows the original ultrasound images with lesion masks, and the bottom row shows the corresponding Grad-CAM heatmaps highlighting the model's attention regions

4 Results and Discussion

Table 1 presents a comparison of our proposed framework against traditional machine learning, deep learning, and multi-modal fusion baselines. Our method

achieves the best overall performance with an accuracy of **90%**, precision of **93%**, and recall of **98%**, indicating strong capability in detecting malignant lesions while minimizing false negatives—a critical requirement in clinical screening. While clinical-only models achieve competitive accuracy, their recall and F1-scores are consistently lower. Image-only deep learning models (e.g., ResNet34/50, EfficientNet) further underperform, particularly in recall, suggesting that imaging alone may lack sufficient discriminative power in challenging cases.

Table 2. Ablation study showing the contribution of each proposed component.

Clinical Knowledge Gate	Uncertainty Attention	Contrastive Learning	Adaptive Focal Loss	Accuracy
-	✓	✓	✓	0.87± 0.03
✓	-	✓	✓	0.82± 0.70
✓	✓	-	✓	0.88± 0.04
✓	✓	✓	-	0.87± 0.05
✓	✓	✓	✓	**0.90 ± 0.07**

Our model also incorporates explicit prediction uncertainty estimation, enhancing its clinical reliability. As shown by average uncertainty scores, clinical features (0.393 ± 0.144) are more stable than image features (0.610 ± 0.094), aligning with clinical intuition—structured information like BI-RADS or lesion margins tends to be more consistent, whereas ultrasound image quality may vary significantly. Figure 3 presents Grad-CAM visualizations for representative benign and malignant lesions. The model's attention aligns with known radiological markers, focusing on irregular margins and heterogeneous echogenicity in malignant cases. This interpretability supports clinical trust and transparency.

To further analyze component contributions, we conducted an ablation study (Table 2). Removing the Clinical Knowledge Gate reduces accuracy by 3%, validating the benefit of domain-aware clinical feature weighting, and excluding the uncertainty-aware attention mechanism results in the most significant drop in accuracy, emphasizing its critical role in robust multi-modal integration.

5 Conclusion

We present an uncertainty-aware cross-modal framework for breast cancer classification that integrates ultrasound imaging with structured clinical data. Our method improves diagnostic accuracy, interpretability, and reliability by combining a domain-informed Clinical Knowledge Gate with uncertainty-guided attention and contrastive cross-modal alignment. Experimental results on the Breast-Lesions-USG dataset demonstrate superior performance over existing baselines, particularly in minimizing false negatives. However, while the dataset provides expert-annotated, biopsy-confirmed labels, its limited size and single-institution

origin constrain generalizability. Future work will focus on multi-center evaluations and larger, more diverse cohorts to ensure broader clinical applicability.

Disclosure of Interests. The authors declare no competing interests relevant to this work.

References

1. AlZoubi, A., Lu, F., Zhu, Y., Ying, T., Ahmed, M., Du, H.: Classification of breast lesions in ultrasound images using deep convolutional neural networks: transfer learning versus automatic architecture design. Med. Biol. Eng. Comput. **62**(1), 135–149 (2024)
2. American cancer society: breast cancer statistics, 2019. CA Cancer J. Clinic. **69**(6), 438–451 (2019)
3. American Cancer Society: Breast Cancer Facts & Figures 2019-2020. American Cancer Society (2021)
4. American College of Radiology: ACR BI-RADS® Atlas, Breast Imaging Reporting and Data System. American College of Radiology, Reston, VA, 5 Edn. (2013). https://www.acr.org/Clinical-Resources/Reporting-and-Data-Systems/Bi-Rads
5. Antropova, N., Huynh, B.Q., Giger, M.L.: A deep feature fusion methodology for breast cancer diagnosis demonstrated on three imaging modality datasets. Med. Phys. **44**(10), 5162–5171 (2017)
6. Begoli, E., Bhattacharya, T., Kusnezov, D.: The need for uncertainty quantification in machine-assisted medical decision making. Nature Mach.Intell. **1**(1), 20–23 (2019)
7. Berg, W.A., Mendelson, E.B., Coscia, J.L., Jr., et al.: Bi-rads ultrasound, pp. 128–173. ACR BI-RADS Atlas, Breast Imaging Reporting and Data System pp (2019)
8. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
9. Cover, T., Hart, P.: Nearest neighbor pattern classification. IEEE Trans. Inf. Theory **13**(1), 21–27 (1967)
10. Han, Z., Wei, B., Zheng, Y., Yin, Y., Li, K., Li, S.: Breast cancer multi-classification from histopathological images with structured deep learning model. Sci. Rep. **7**(1), 4172 (2017)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
12. Hooley, R.J., Greenberg, K.L., Stackhouse, R.M., Geisel, J.L., Butler, R.S., Philpotts, L.E.: Screening us in patients with mammographically dense breasts: initial experience with connecticut public act 09–41. Radiology **265**(1), 59–69 (2013)
13. Hosmer Jr, D.W., Lemeshow, S., Sturdivant, R.X.: Applied Logistic Regression, vol. 398. John Wiley & Sons (2013)
14. Huang, S.C., Pareek, A., Seyyedi, S., Banerjee, I., Lungren, M.P.: Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. NPJ Digit. Med. **2**(1), 1–9 (2019)
15. Islam, M.R., et al.: Enhancing breast cancer segmentation and classification: an ensemble deep convolutional neural network and u-net approach on ultrasound images. Mach. Learn. Appl. **16**, 100555 (2024)
16. Lee, S.H., et al.: Differentiation of benign from malignant solid breast masses: comparison of two-dimensional and three-dimensional shear-wave elastography. Eur. Radiol. **23**(4), 1015–1026 (2013)

17. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2980–2988 (2017)
18. Lei, X., et al.: Deep learning in medical ultrasound analysis: a review. Engineering **5**(2), 261–275 (2019)
19. Lu, J., Batra, D., Parikh, D., Lee, S.: VilBERT: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In: Advances in Neural Information Processing Systems, vol. 32, pp. 13–23 (2019)
20. Mendelson, E.B., Bohm-Velez, M., Berg, W.A., Whitman, G.J., Feldman, M.K., Madjar, H.: Toward a standardized breast ultrasound lexicon, bi-rads ultrasound. Semin. Roentgenol. **48**(4), 271–288 (2013)
21. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
22. Pawłowska, A., et al.: Curated benchmark dataset for ultrasound based breast lesion analysis. Sci. Data **11**(1), 148 (2024)
23. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748–8763. PMLR (2021)
24. Raza, A., Ullah, N., Khan, J.A., Assam, M., Guzzo, A., Aljuaid, H.: DeepBreastCancerNet: a novel deep learning model for breast cancer detection using ultrasound images. Appl. Sci. **13**(4), 2082 (2023)
25. Shen, D., Wu, G., Suk, H.I.: Deep learning in medical image analysis. Annu. Rev. Biomed. Eng. **21**, 221–248 (2019)
26. Siegel, R.L., Miller, K.D., Wagle, N.S., Jemal, A.: Cancer statistics, 2023. CA Cancer J. Clin. **73**(3), 233–254 (2023)
27. Stavros, A.T., Thickman, D., Rapp, C.L., Dennis, M.A., Parker, S.H., Sisney, G.A.: Solid breast nodules: use of sonography to distinguish between benign and malignant lesions. Radiology **196**(1), 123–134 (1995)
28. Tan, M., Le, Q.: EfficientNet: rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning, pp. 6105–6114. PMLR (2019)
29. Wang, D., Khosla, A., Gargeya, R., Irshad, H., Beck, A.H.: Deep learning for identifying metastatic breast cancer. Nature **529**(7587), 484–488 (2016)
30. Yala, A., Lehman, C., Schuster, T., Portnoi, T., Barzilay, R.: A deep learning mammography-based model for improved breast cancer risk prediction. Radiology **292**(1), 60–66 (2019)
31. Zhang, T., et al.: Radiomics and artificial intelligence in breast imaging: a survey. Artif. Intell. Rev. **56**, 857–892 (2023). https://doi.org/10.1007/s10462-023-10543-y



Comparing Conditional Diffusion Models for Synthesizing Contrast-Enhanced Breast MRI from Pre-contrast Images

Sebastian Ibarra[1(✉)], Javier del Riego[2], Alessandro Catanese[3], Julian Cuba[4], Julian Cardona[4], Nataly Leon[4], Jonathan Infante[4], Karim Lekadir[1,5], Oliver Diaz[1,6], and Richard Osuala[1]

[1] Departament de Matemàtiques i Informàtica, Universitat de Barcelona, Barcelona, Spain
{sebastian.ibarra,richard.osuala}@ub.edu
[2] Parc Taulí Hospital Universitari, Sabadell, Spain
[3] Hospital Germans Trias i Pujol, Barcelona, Spain
[4] Siemens Healthineers, America LAM, San Juan, USA
[5] Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain
[6] Computer Vision Center, Bellaterra, Spain

Abstract. Dynamic contrast-enhanced (DCE) MRI is essential for breast cancer diagnosis and treatment. However, its reliance on contrast agents introduces safety concerns, contraindications, increased cost, and workflow complexity. To this end, we present pre-contrast conditioned denoising diffusion probabilistic models to synthesize DCE-MRI, introducing, evaluating, and comparing a total of 22 generative model variants in both single-breast and full breast settings. Towards enhancing lesion fidelity, we introduce both tumor-aware loss functions and explicit tumor segmentation mask conditioning. Using a public multicenter dataset and comparing to respective pre-contrast baselines, we observe that subtraction-image-based models consistently outperform post-contrast-based models across five complementary evaluation metrics. Apart from assessing the entire image, we also separately evaluate the region of interest, where both tumor-aware losses and segmentation mask inputs improve evaluation metrics. The latter notably enhance qualitative results capturing contrast uptake, albeit assuming access to tumor localization inputs that are not guaranteed to be available in screening settings. A reader study involving 2 radiologists and 4 MRI technologists confirms the high realism of the synthetic images, indicating an emerging clinical potential of generative contrast-enhancement. We share our codebase at https://github.com/sebastibar/conditional-diffusion-breast-MRI.

Keywords: Contrast Agent · Generative · DCE-MRI · Breast Cancer

1 Introduction

Breast cancer is the most frequently diagnosed cancer in women, with 2.3 million new cases and 670,000 deaths in 2022 [1]. Early detection improves outcomes,

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
T. Zhang et al. (Eds.): Deep-Breath 2025, LNCS 16142, pp. 226–236, 2026.
https://doi.org/10.1007/978-3-032-05559-0_23

and breast magnetic resonance imaging (MRI) offers high sensitivity for screening high-risk individuals [12]. In this regard, dynamic contrast-enhanced MRI (DCE-MRI) is the clinical gold standard, leveraging gadolinium-based contrast agents (GBCAs) to localize and visualize tumor vascularity [14]. Despite its high diagnostic value, DCE-MRI is time-consuming and expensive as well as contraindicated in vulnerable populations [2,18]. GBCAs further pose risks such as nephrogenic systemic fibrosis [22], gadolinium retention [27], allergic-like reactions as well as other Symptoms Associated with Gadolinium Exposure (SAGE) [15]. These limitations have driven interest in contrast-free imaging alternatives. Deep generative models–particularly conditional architectures–can synthesize post-contrast images from non-contrast inputs, with promising early results [3,18,19,23,29]. With other generative modeling techniques often suffering from instability, low generalization, or poor perceptual quality [4], image-level denoising diffusion probabilistic models [8] (DDPMs) offer a compelling alternative via iterative denoising, enabling stable, high-fidelity synthesis [24]. Recent studies demonstrate their effectiveness in medical image denoising and anomaly detection, as well contrast-agent dose reduction [16], and segmentation-guided anatomy generation in breast MRI [9]. Although latent diffusion models [20] have been recently introduced for pre- to post-contrast breast MRI synthesis [19], their image-level counterparts are still an underexplored method in this domain, e.g., missing comprehensive benchmarking across multiple model variants and preprocessing methods, supervision strategies, and clinical settings. While DDPMs have recently been applied to breast MRI synthesis [16], we present the first quantitative benchmark of 22 conditional variants, including single-breast and ROI-aware models on a large-scale multicenter dataset, along with novel insights from an expert reader study, summarized as the following contributions:

- Implementation and benchmarking framework for pre-contrast conditioned DDPMs for DCE-MRI synthesis.
- Proposal of tumor-aware losses and segmentation-guided model conditioning.
- Comparison between post-contrast and subtraction image targets, as well as full and single-breast synthesis strategies.
- Inclusion of an expert-reader study alongside quantitative and qualitative lesion and image-level assessment, noting synthesis realism, particularly in subtraction-based and tumor-aware models.

2 Material and Methods

2.1 Dataset and Preprocessing

We use the MAMA-MIA dataset [5], a large-scale benchmark aggregating 1,506 pre-treatment DCE-MRI cases from four public datasets across multiple institutions, vendors, and imaging protocols. This heterogeneity supports robust assessment of model generalizability for tasks such as segmentation, treatment response prediction, and contrast synthesis. It includes lesion-type annotations, comprising luminal A (43.6%), luminal B (19.5%), HER2-enriched (14.4%), and triple-negative breast cancer (22.5%) cases.

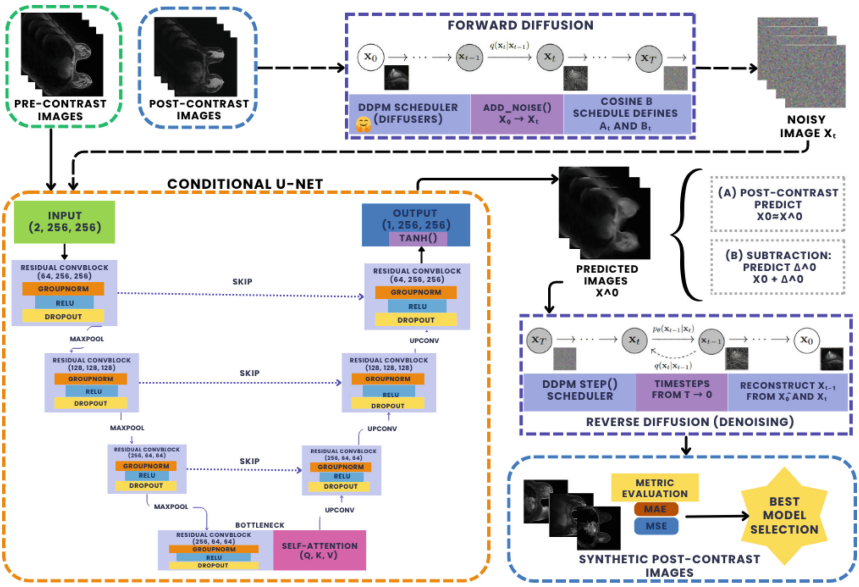


Fig. 1. Overview depicting both post-contrast and subtraction-based diffusion model training pipelines and model architectures

For each patient, we selected Phase 0 (pre-contrast) as model input and Phase 1 (early post-contrast) as target, ensuring anatomical alignment and temporal relevance. Based on MAMA-MIA's expert-corrected tumor segmentation masks, all slices containing tumor voxels were extracted, along with 20% of adjacent non-tumor slices to increase data diversity. For slice extraction, volumes were loaded as 3D NIfTI files, sliced axially, and then, each 2D axial slice was min-max normalized to [0,1] based on its own intensity range, then scaled to [0,255] and exported as 8-bit PNG. Data were split into four subsets (unilateral/bilateral, train/test), with accompanying metadata including patient ID, slice index, and tumor label. Patients were classified as unilateral or bilateral based on metadata; all images retain the full axial field of view. The final dataset contains 92,838 paired 2D images across contrast phases with a split of 80% train and 20% test. A visualization of the preprocessing pipeline is depicted in the Supplementary Material of our study.

2.2 Contrast-Enhancing Generative Modeling

As illustrated in Fig. 1, we develop and compare two conditional DDPMs for synthetic post-contrast MRI generation. Firstly, the post-contrast DDPM (PC) directly predicts the full post-contrast image from a respective pre-contrast input image. A second alternative subtraction-based DDPM (SUB) learns to predict the residual enhancement ($x_{\text{post}} - x_{\text{pre}}$), which is added back to the pre-contrast

image to reconstruct the corresponding post-contrast image. Both models were implemented in full-breast and single-breast variants.

Post-Contrast Conditional DDPM (PC): To this end, we implement a conditional U-Net model with residual blocks and a bottleneck self-attention layer which takes a 2-channel input (pre-contrast + noisy post-contrast at timestep t) and predicts the post-contrast image $\hat{x}_0$. Following the image-predictive variant in [25], the model regresses $\hat{x}_0$ instead of noise ϵ, enforcing pixel-level consistency and improving anatomical fidelity. Gaussian noise is added to x_0 using a cosine schedule. With c as the conditioning pre-contrast image, during denoising the model computes the reverse posterior mean according to:

$$\mu_\theta(x_t, c, t) = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}} \hat{x}_0(x_t, c, t) \right)$$

We further define our training objective as a weighted combination of image-level comparison loss functions based on the intuition of a balance of pixel accuracy, structure, and spatial smoothness: $\mathcal{L}_{\text{total}} = 0.3\,\mathcal{L}_{\text{MAE}} + 0.6\,\mathcal{L}_{\text{Perceptual}} + 0.15\,\mathcal{L}_{\text{Total Variation (TV) [21]}} + 0.05\,\mathcal{L}_{\text{MSE}}$. To improve convergence efficiency, we adopt the AdamW optimizer [13], exponential moving average (EMA) models with $\lambda = 0.999$ and mixed-precision training using [6].

Subtraction-Based Conditional DDPM (SUB): This variant learns to predict the scaled subtraction image $x_{\text{sub}} = (x_{\text{post}} - x_{\text{pre}})/0.5$, which represents the localized contrast enhancement. During inference, the synthetic post-contrast image is reconstructed as $x_{\text{post}}^{\text{gen}} = 0.5 \cdot \hat{x}_0 + x_{\text{pre}}$, where $\hat{x}_0$ denotes the predicted clean subtraction image. The architecture, noise schedule, and training setup mirror those of the post-contrast DDPM. However, the loss is computed on the reconstructed post-contrast image to ensure anatomical fidelity and contrast realism.

2.3 Tumor-Aware Model Training

We implement two tumor region-of-interest (ROI) aware models. In the first, the expert segmentation mask is concatenated as an additional channel to the input, thereby facilitating the model to focus on learning contrast enhancement patterns within lesion-containing regions of the pre-contrast images. Instead of modifying the model input, the second ROI-aware variant receives the full pre-contrast image but introduces lesion-specific supervision through a spatially-weighted loss guided by the expert segmentation mask M. Based on M, the ROI loss $\mathcal{L}_{\text{ROI}}$ combines (1) pixel-wise MAE and MSE within the tumor region, (2) a VGG-based perceptual loss over the tumor region, and (3) a total variation (TV) [21] term to penalize high-frequency noise and to promote spatial smoothness.

Motivated by the intuition of designing a well-rounded set of complementary region-aware losses, and supported by visual observations of underenhancement in tumor regions, we add a contrast-specific MAE term that penalizes only underestimated signals using ReLU-masked residuals: $\mathcal{L}_{\text{MAEcontrast}} = \text{MAE}([\hat{x}_{\text{ROI}} -$

$x_{\text{pre}} \cdot M]_+, [x_{\text{ROI}} - x_{\text{pre}} \cdot M]_+)$. To preserve mean enhancement intensity, we also include an intensity loss term defined as $\mathcal{L}_{\text{intensity}} = |\mu(\hat{x}_{\text{ROI}}) - \mu(x_{\text{ROI}})|$. The final loss is a weighted sum: $\mathcal{L}_{\text{total}} = 0.3\,\mathcal{L}_{\text{global}} + 0.6\,\mathcal{L}_{\text{ROI}} + 0.05\,\mathcal{L}_{\text{MAEcontrast}} + 0.05\,\mathcal{L}_{\text{intensity}}$, in order to balance tumor-aware fidelity with global accuracy via the $\mathcal{L}_{\text{global}}$ loss computed on the full image as defined in 2.2.

2.4 Reader Study: Clinical Validation

Six domain experts evaluated the synthesis realism in a three-part visual assessment, including two radiologists with 11+ and 9+ years practicing experience and four expert MRI technologists with 10 to 15+ years of experience. In Task 1 (Discrimination), 15 mixed images (10 synthetic, 5 real) are shown individually to assess distinguishability between real and synthetic. In Task 2 (Comparative), readers were shown randomly selected triplets (pre-contrast, real post, synthetic post) in order to identify the real image between the two post-contrast images. In Task 3 (Annotation), experts annotate differences and score realism within triplets already labeled as real and synthetic. A semi-structured follow-up discussion captured further insights on diagnostic relevance and clinical integration.These qualitative assessments related to tumor conspicuity and enhancement realism, provided insights into diagnostic fidelity.

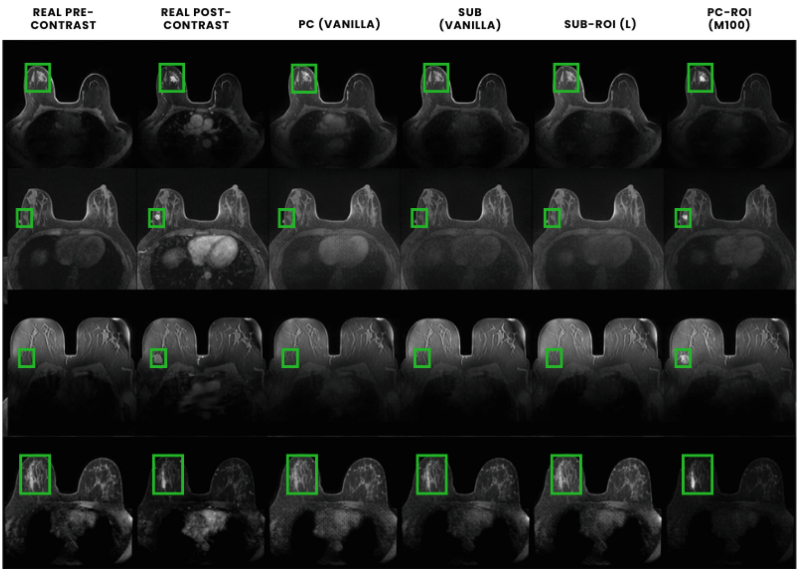


Fig. 2. Qualitative synthetic test set samples comparing model variants with real pre- (col 1) and post-contrast (col 2) images. The variants include the vanilla post-contrast DDPM (col 3), the vanilla subtraction DDPM (col 4), ROI-aware loss subtraction DDPM (col 5), and ROI-mask conditioned post-contrast DDPM (col 6). The lesion area is highlighted using green bounding boxes (Color figure online)

3 Results

3.1 Qualitative Evaluation and Comparison

Figure 2 presents a visual comparison across model variants for four patient cases. While all models exhibit high-quality synthesis and realistic appearance, some limitations remain in accurately detecting lesions and rendering appropriate contrast. Compared to the vanilla post-contrast (PC) and subtraction (SUB) DDPMs, ROI-aware models, in particular, mask-conditioned models such as PC-ROI$_{(M100)}$ yield visibly improved enhancement within tumor regions with promising synthetic contrast patterns. Such mask-guided approaches are particularly relevant in applications with known approximate tumor locations, which include treatment response assessment, longitudinal surveillance or pre-treatment monitoring. Training on full-breast images allows to learn patterns concerning anatomical symmetry relevant to tumor characterization. In contrast, single-breast models, while potentially benefiting from an enhanced focus on a single breast, on the other hand, can miss contralateral cues, which can be relevant in complex contrast enhancement and lesion detection cases. For qualitative single breast results, we refer the reader to the Supplementary Material of our study Notably, ROI-aware supervision commonly enhances boundary sharpness and better approximates the lesion intensity observed in real DCE-MRI scans. Overall, subtraction-based and ROI-guided variants exhibit stronger lesion reconstruction quality.

3.2 Quantitative Assessment and Benchmarking

We quantitatively evaluate our model variants calculating image synthesis evaluation metrics for both full image and for the tumor-containing region-of-interest (ROI) of each image. Apart from benchmarking PC and SUB image synthesis, we also comparatively assess full breast MRI and single-breast image synthesis depicted in Table 1 and Table 2, respectively. To this end, we apply mutually-complementary [18] image quality metrics, namely, Mean Absolute Error (MAE), Structural Similarity Index (SSIM) [26], Peak Signal-to-Noise Ratio (PSNR), Learned Perceptual Image Patch Similarity (LPIPS) [28], Fréchet Inception Distance (FID) [7] and Fréchet Radiomics Distance (FRD) [10,19]. The results demonstrate the ability of diffusion models to successfully synthesize both PC and SUB images from pre-contrast inputs in both full-breast and single-breast settings. When comparing vanilla models for PC and SUB image synthesis while considering their respective *real pre vs real target* baselines, particularly the SUB$_{(Vanilla)}$ model improves considerably over its baseline to this end outperforming its PC counterpart. This observation holds true for both full-image as well as ROI-based metric computation across distribution-based and image-level metrics. This trend of SUB outperforming PC synthesis is similarly observable for both full-breast MRI slice models and cropped single-breast image synthesis models. Comparing such subtraction based single and full-breast models, the full-breast models more consistently outperformed the respective *real pre vs real*

Table 1. Quantitative results and ablations for *full breast DCE-MRI synthesis* based on distribution-level metrics (FRD↓, FID↓) as well as paired image comparison metrics (MAE↓, LPIPS↓, SSIM↑, PSNR↑). Prefixes: PC=Post-Contrast, SUB = Subtraction. Suffixes: (100) = Model trained for 100 epochs rather than 50 epochs, (M) = Region-of-Interes(ROI)-masked, (L) = ROI-aware loss. PC models are compared to corresponding *Real PC* and SUB models to corresponding *Real SUB* images. Best values in **bold**

Model Variant	FRD↓	FID↓	MAE↓	LPIPS↓	SSIM↑	PSNR↑
Full-Breast (metrics computed on full image)						
Real Pre vs Real PC	**11.02**	**57.79**	.062 ± .206	**.193 ± .036**	.811 ± .021	23.09 ± 2.86
$PC_{(Vanilla)}$	59.10	74.49	.072 ± .031	.203 ± .033	.815 ± .066	22.10 ± 3.51
$PC_{(Vanilla100)}$	58.85	72.56	.071 ± .031	.202 ± .033	.816 ± .066	22.24 ± 3.55
$PC\text{-}ROI_{(M)}$	29.29	73.29	.064 ± .026	.203 ± .032	.823 ± .063	23.30 ± 3.62
$PC\text{-}ROI_{(M100)}$	31.23	72.68	**.060 ± .023**	.197 ± .032	**.826 ± .063**	**23.94 ± 3.56**
$PC\text{-}ROI_{(L)}$	23.45	75.13	.065 ± .022	.235 ± .043	.819 ± .060	22.85 ± 3.01
Real Pre vs Real SUB	100.46	165.91	.862 ± .055	.544 ± .034	.028 ± .024	7.05 ± .45
$SUB_{(Vanilla)}$	28.99	69.75	**.054 ± .021**	**.194 ± .037**	**.847 ± .061**	**25.63 ± 3.35**
$SUB\text{-}ROI_{(L)}$	**21.29**	**67.07**	**.054 ± .017**	.195 ± .035	.826 ± .062	24.11 ± 2.75
Full-Breast (metrics computed on ROI)						
Real Pre vs Real PC	42.10	95.64	.209 ± .054	**.140 ± .061**	**.495 ± .191**	16.94 ± 3.38
$PC_{(Vanilla)}$	59.10	129.10	.334 ± .134	.310 ± .098	.303 ± .193	14.33 ± 3.49
$PC_{(Vanilla100)}$	68.85	140.69	.389 ± .151	.324 ± .097	.283 ± .186	13.17 ± 3.41
$PC\text{-}ROI_{(M)}$	29.29	70.69	.185 ± .066	.221 ± .084	.458 ± .182	19.01 ± 2.77
$PC\text{-}ROI_{(M100)}$	27.82	**67.04**	**.175 ± .057**	.215 ± .082	.475 ± .171	**19.28 ± 2.38**
$PC\text{-}ROI_{(L)}$	**21.29**	79.06	.265 ± .104	.258 ± .090	.302 ± .182	16.34 ± 3.02
Real Pre vs Real SUB	211.29	210.50	.391 ± .052	.388 ± .063	.326 ± .207	8.66 ± 3.24
$SUB_{(Vanilla)}$	29.44	98.37	.301 ± .133	.281 ± .095	.342 ± .207	15.32 ± 3.67
$SUB\text{-}ROI_{(L)}$	**23.45**	**89.48**	**.247 ± .107**	**.265 ± .094**	**.374 ± .210**	**16.84 ± 3.42**

SUB baseline across metrics than their single-breast counterparts for across both synthetic ROI and full synthetic image evaluation metrics. For full breast MRI slice generation, the inclusion of a tumor-aware loss (e.g. in $SUB\text{-}ROI_{(L)}$) generally further improved the synthesis performance for ROI-based metrics, while only resulting in improvements on distribution leverl (FRD, FID) for full image metrics (compared to $SUB_{(Vanilla)}$). In single breast generation models, the same trend is observable for both ROI-based metrics and full-image metrics, with the exception of LPIPS in the latter, where single-breast $SUB\text{-}ROI_{(L)}$ (.249 ± .052) outperforms $SUB_{(Vanilla)}$ (.266 ± .061).

3.3 Reader Study: Results and Evaluation

For reader assessment, all three tasks were based on outputs of ($SUB_{(Vanilla)}$) due to its promising performance, with additional comparison of ROI-aware and ROI-masked model variants. Task 1 showed moderate difficulty, with readers

Table 2. Quantitative results and ablations for single-breast image synthesis based on the same metrics and prefixes as in Table 1

Model Variant	FRD↓	FID↓	MAE↓	LPIPS↓	SSIM↑	PSNR↑
Single-Breast (metrics computed on full image)						
Real Pre vs Real PC	**11.02**	**79.89**	.120 ± .053	**.257 ± .057**	.634 ± .126	**20.82 ± 3.09**
$PC_{(Vanilla)}$	177.10	82.29	.128 ± .047	.294 ± .060	.700 ± .097	18.87 ± 2.87
$PC\text{-}ROI_{(L)}$	60.74	104.03	**.111 ± .041**	.304 ± .068	**.702 ± .096**	20.34 ± 3.02
Real Pre vs Real SUB	107.62	131.16	.736 ± .101	.514 ± .030	.047 ± .054	8.12 ± 1.01
$SUB_{(Vanilla)}$	59.82	88.10	**.092 ± .036**	.266 ± .061	**.715 ± .095**	**22.52 ± 2.80**
$SUB\text{-}ROI_{(L)}$	**43.55**	**80.47**	.105 ± .045	**.249 ± .052**	.695 ± .099	21.89 ± 2.97
Single-Breast (metrics computed on ROI)						
Real Pre vs Real PC	71.02	**79.89**	**.126 ± .154**	**.257 ± .057**	**.634 ± .126**	**20.82 ± 3.09**
$PC_{(Vanilla)}$	177.10	210.50	.575 ± .227	.499 ± .092	.172 ± .167	10.17 ± 3.39
$PC\text{-}ROI_{(L)}$	**60.74**	151.72	.340 ± .138	.438 ± .104	.187 ± .181	14.31 ± 3.13
Real Pre vs Real SUB	85.35	**101.60**	.321 ± .107	.368 ± .081	**.288 ± .235**	13.68 ± 3.44
$SUB_{(Vanilla)}$	59.82	136.59	.307 ± .161	.402 ± .109	.256 ± .207	15.44 ± 3.95
$SUB\text{-}ROI_{(L)}$	**43.55**	104.60	**.266 ± .125**	**.366 ± .101**	.278 ± .205	**16.53 ± 3.44**

achieving between 60% and 80% accuracy. In Task 2, they consistently identified the real post-contrast image, noting improvement due to feedback from Task 1. Realism scores in Task 3 ranged from 5 to 7 out of 10. Annotations highlighted strong anatomical fidelity in synthetic images, particularly in breast parenchyma, skin boundaries, and mammary tissue. Several experts praised artifact simulation and the preservation of coarse contrast in ROI-guided models. However, limitations were also noted, including insufficient tumor enhancement (particularly in early contrast phases), flat background textures, grid-like artifacts, and asymmetries in the heart and chest regions. These were identified as indicators of synthetic images. By uncovering these aspects, our study establishes a reference point for qualitative evaluation and provides actionable insights to guide the development of future models. The readers emphasized the clinical potential of the generative modeling approach and viewed the ROI-aware and ROI-masked variants as promising for specific tasks such as treatment monitoring, response assessment, or as adjunct tools in abbreviated MRI protocols, where full contrast dynamics can be unavailable or undesired. Selected expert annotations are visualized in the Supplementary Material.

4 Discussion and Conclusions

We proposed and tested pre-contrast-conditioned DDPM frameworks where our subtraction-based models ($SUB_{(Vanilla)}$) quantitatively outperformed the post-contrast ones ($PC_{(Vanilla)}$) based on respective baseline comparisons. In subtraction variants the synthesis task is likely simplified by allowing the model to focus exclusively on the residual of the contrast signal, while anatomy remains

mostly unchanged. ROI losses improved quantitative results without requiring segmentation masks at inference, thus enabling lesion detection applications. For scenarios where lesion localization is known before, e.g. in cancer treatment, segmentation mask conditioned models notably improve qualitative ROI results. This design reflects our goal to evaluate both conditioning strategies, rather than selecting a single best-performing model. Expert reader evaluation confirmed anatomical realism, especially for subtraction-based and ROI-conditioned models. While artifacts as limited tumor contrast, flat backgrounds, or grid patterns were observed, ROI-guided variants were perceived as more clinically promising, particularly for monitoring or abbreviated protocols. We acknowledge the sample image size, which limits statistical power and warrants broader clinical validation of these valuable insights.

Future work will explore ablations of loss components, loss on subtraction images, and optimization of loss weightings. Extensions to 3D synthesis, DCE-MRI temporal modeling [11,19,23], and comparisons with latent diffusion and GAN models [16,17] are also planned. Overall, our promising results motivate further exploration of contrast-free imaging as a safe, scalable alternative for patients where contrast agents are contraindicated

Disclosure of Interests. The authors declare no conflict of interest. Some authors are affiliated with Siemens Healthineers. This work reflects their personal scientific views and does not represent the company's official position or endorsement.

References

1. Bray, F., et al.: Global cancer statistics 2022: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: Can. J. Clin. **74**(3), 229–263 (2024)
2. Chhor, C.M., Mercado, C.L.: Abbreviated mri protocols: wave of the future for breast cancer screening. Am. J. Roentgenol. **208**(2), 284–289 (2017)
3. Denck, J., Guehring, J., Maier, A., Rothgang, E.: Enhanced magnetic resonance image synthesis with contrast-aware generative adversarial networks. J. Imaging **7**(8), 133 (2021)
4. Dimitriadis, A., Trivizakis, E., Papanikolaou, N., Tsiknakis, M., Marias, K.: Enhancing cancer differentiation with synthetic mri examinations via generative models: a systematic review. Insights Imaging **13**(1), 188 (2022)
5. Garrucho, L., et al.: A large-scale multicenter breast cancer dce-mri benchmark dataset with expert segmentations. Sci. Data **12**(1), 453 (2025). https://doi.org/10.1038/s41597-025-04707-4
6. Gugger, S., et al.: Accelerate: training and inference at scale made simple, efficient and adaptable. (2022), https://github.com/huggingface/accelerate
7. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
8. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Adv. Neural. Inf. Process. Syst. **33**, 6840–6851 (2020)

9. Konz, N., Chen, Y., Dong, H., Mazurowski, M.A.: Anatomically-controllable medical image generation with segmentation-guided diffusion models. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 88–98. Springer (2024)
10. Konz, N., et al.: Fréchet radiomic distance (frd): a versatile metric for comparing medical imaging datasets (2025), https://arxiv.org/abs/2412.01496
11. Lang, D.M., Osuala, R., Spieker, V., Lekadir, K., Braren, R., Schnabel, J.A.: Temporal neural cellular automata: application to modeling of contrast enhancement in breast MRI. arXiv preprint arXiv:2506.18720, https://arxiv.org/abs/2506.18720 (2025)
12. Lehman, C.D., et al.: Screening women at high risk for breast cancer with mammography and magnetic resonance imaging. Cancer **103**(9), 1898–1905 (2005)
13. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
14. Mann, R.M., Cho, N., Moy, L.: Breast mri: state of the art. Radiology **292**(3), 520–536 (2019)
15. McDonald, R.J., Weinreb, J.C., Davenport, M.S.: Symptoms associated with gadolinium exposure (sage): a suggested term. Radiology **302**(2), 270–273 (2022)
16. Müller-Franzes, G., et al.: Diffusion probabilistic versus generative adversarial models to reduce contrast agent dose in breast mri. Eur. Radiol. Exp. **8**(1), 53 (2024)
17. Osuala, R., et al.: Simulating dynamic tumor contrast enhancement in breast mri using conditional generative adversarial networks. arXiv preprint arXiv:2409.18872 (2024)
18. Osuala, R., et al.: Simulating dynamic tumor contrast enhancement in breast mri using conditional generative adversarial networks. J. Med. Imaging **12**(S2), S22014–S22014 (2025)
19. Osuala, R., et al.: Towards learning contrast kinetics with multi-condition latent diffusion models. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 713–723. Springer (2024)
20. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10684–10695 (2022)
21. Rudin, L.I., Osher, S., Fatemi, E.: Nonlinear total variation based noise removal algorithms. Physica D **60**(1–4), 259–268 (1992)
22. Sardanelli, F., Cozzi, A., Trimboli, R.M., Schiaffino, S.: Gadolinium retention and breast mri screening: more harm than good? Am. J. Roentgenol. **214**(2), 324–327 (2020)
23. Schreiter, H., et al.: Virtual dynamic contrast enhanced breast mri using 2d u-net architectures. In: Deep Breast Workshop on AI and Imaging for Diagnostic and Treatment Challenges in Breast Care, pp. 85–95. Springer (2024)
24. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456 (2021)
25. Von Platen, P., Patil, S., Cuenca, P., et al.: Diffusers: state-of-the-art diffusion models for image and audio generation in pytorch. https://github.com/huggingface/diffusers (2022)
26. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004)
27. Xie, H., et al.: Magnetic resonance imaging contrast enhancement synthesis using cascade networks with local supervision. Med. Phys. **49**(5), 3278–3287 (2022)

28. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 586–595 (2018)
29. Zhang, T., et al.: Synthesis of contrast-enhanced breast mri using t1-and multi-b-value dwi-based hierarchical fusion network with attention mechanism. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 79–88. Springer (2023)



Adapting Vision Language Models for Structured Clinical Description Generation in Mammography

André C. Castro[1(✉)], Dayane Rodrigues[1], and Arlindo R. G. Filho[2]

[1] Institute of Informatics, Federal University of Goiás, Goiânia, Brazil
andre.castro@discente.ufg.br

[2] Advanced Knowledge Center for Immersive Technologies, Goiânia, Brazil

Abstract. Automated report generation for mammographic images remains a significant challenge in clinical radiology, particularly in resource-constrained environments where specialized artificial intelligence (AI) infrastructure is limited. This work evaluates open-access vision-language models (VLMs) and fine-tunes the best-performing one to generate descriptions of mammographic images. We compare four baseline models: BLIP, MiniGPT-4, GIT, and CLIP+GPT-2, using a multi-agent LLM evaluation protocol with GPT-4o mini, GPT-4.1 mini, and Gemini 2.5 Flash. Our experiments on the VinDr-Mammo dataset show that CLIP+GPT-2 achieves the highest normalized clinical relevance score (0.4687). After fine-tuning on 15,000 training images, the model demonstrates improvements: BLEU scores increased by 3282.6%, ROUGE-L by 2055.3%, and METEOR by 1737.2%. These results demonstrate that accessible and computationally efficient vision-language models can be effectively adapted for structured clinical description generation, offering a practical solution for deploying AI in mammography screening programs.

Keywords: Medical Imaging · Mammography · Vision-Language Models · Deep Learning · Computer-Aided Diagnosis · Structured Description Generation

1 Introduction

Mammographic screening is fundamental for early breast cancer detection, with radiologists interpreting millions of examinations annually worldwide [2]. However, the increasing volume of screening programs and the shortage of specialized radiologists create significant bottlenecks in healthcare delivery [1,3]. Automated report generation systems could substantially alleviate this burden by providing preliminary assessments and structured descriptions of mammographic findings [4].

Recent advances in vision-language models (VLMs) have shown remarkable progress in generating natural language descriptions from images across general domains [5,6]. However, their application to medical imaging, particularly mammography, faces unique challenges including domain specificity, clinical terminology requirements, and the critical need for accurate interpretation of subtle radiological findings [7]. While large-scale models like GPT-4V demonstrate impressive

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
T. Zhang et al. (Eds.): Deep-Breath 2025, LNCS 16142, pp. 237–247, 2026.
https://doi.org/10.1007/978-3-032-05559-0_24

performance [8], their computational requirements and proprietary nature limit accessibility in clinical environments with constrained resources [9].

This work addresses the critical need for accessible solutions by investigating whether open-access vision-language models can be effectively adapted for mammographic report generation. Our main contributions are: (1) A comprehensive evaluation framework using multi-agent LLM scoring to assess clinical relevance without human annotation; (2) Systematic comparison of four open-access VLMs for structured clinical description generation on mammographic images (3); A successful fine-tuning methodology that improves performance using structured clinical annotations; (4) Demonstration that accessible models can achieve clinically meaningful results suitable for deployment in resource-constrained environments.

2 Related Work

Medical image-to-text generation has gained increasing attention, particularly in the context of radiological report generation. Early methods employed encoder-decoder architectures with attention mechanisms applied to general medical datasets [10–12]. More recent approaches leverage transformer-based models for chest X-ray interpretation, including cross-modal memory networks [13] and hierarchical alignment strategies [14]. Efforts have also focused on improving factual accuracy in generated reports [15].

In mammography, most prior work has emphasized classification or lesion detection [16–19] with limited progress on automated generation of structured clinical descriptions. Notable contributions include deep learning methods for breast cancer screening [20], and breast density estimation [21]. However, full-sentence generation for mammographic findings remains underexplored.

The advent of large vision-language models (VLMs) like GPT-4V [8] has sparked interest in foundation models for medical imaging [22,23]. Models such as GLoRIA [24], PaLM-E [25], and InstructBLIP [26] highlight the potential of multimodal instruction-tuned systems. Nonetheless, most vision-language models (VLMs) remain closed-source and inaccessible, limiting their practical deployment. Our work addresses this gap by systematically evaluating and fine-tuning open-access models for mammographic report generation in real-world, resource-constrained settings.

3 Methodology

3.1 Model Selection Criteria

To enable practical deployment in real-world clinical environments, we focused on open-access vision-language models (VLMs) that can be fine-tuned and evaluated using a single NVIDIA V100 GPU (32GB VRAM) on a DGX platform. Our goal was to evaluate pre-trained image captioning models that can generate descriptive outputs for mammograms without requiring specialized infrastructure or extensive memory.

3.2 Models Evaluated

We selected four models as baselines for evaluation, considering inference cost, model size, and ease of integration.

BLIP (Salesforce/blip-image-captioning-base): A transformer-based image captioning model trained on general image-text datasets such as COCO [5]. It is fast and lightweight, capable of generating generic yet coherent descriptions even in the medical imaging domain.

MiniGPT-4: Built on the BLIP-2 architecture, this model uses a visual encoder paired with a Flan-T5 language model decoder, trained on multimodal instruction-following data [27]. It offers a favorable trade-off between performance and computational cost.

GIT (microsoft/git-base-coco): A compact generative transformer for image captioning [28]. Despite its efficiency, initial results on mammograms showed it tends to produce overly generic or unrelated outputs.

CLIP + GPT-2: A modular pipeline combining the CLIP image encoder [6] with GPT-2 for caption generation. Descriptions are driven by semantic similarity between the image and a curated list of candidate phrases. This approach showed surprising robustness and clinical relevance given its simplicity.

These models were chosen to assess whether open-access, readily deployable vision-language architectures can generate clinically meaningful descriptions for mammographic images, serving as potential assistive tools in diagnostic workflows, especially in settings with limited access to proprietary AI systems.

3.3 Label Description Generation

To create clinically informed references for evaluation, we implemented a rule-based description generator using structured annotations from the VinDr-Mammo dataset [29], which contains 20,000 mammographic images. The dataset provides two key annotation files:

`breast-level_annotations.csv` and `finding_annotations.csv`, which include breast-level BI-RADS, density assessments, and localized radiological findings per image.

The VinDr-Mammo dataset, comprising images from two distinct hospitals across a three-year span (20182020), enhances the potential for model generalization.

A single-view approach was adopted to reduce the complexity of correlating bilateral images. While it limits asymmetry assessment, this choice ensures computational simplicity and feasibility.

Our description generator maps structured fields into English natural language using a set of controlled templates:

- **Findings** such as *Mass*, *Calcification*, and *Architectural Distortion* are mapped to phrases like *Mass identified* and *Calcifications present*.
- **Breast density** levels (A–D) are described using the ACR BI-RADS lexicon, such as *Breast tissue is heterogeneously dense*.

– **BI-RADS categories** (0–6) are translated into full assessment statements with the suffix *(AI suggestion – not diagnostic)* to reinforce the non-diagnostic nature.

Each image-level description is assembled by combining these elements with standard prefix and suffix phrases. A typical example is:

AI-assisted analysis suggests: Mass identified. Breast tissue is heterogeneously dense. Suspicious finding requiring histologic investigation (AI suggestion - not diagnostic). Radiologist review required for final diagnosis.

When no localized findings are available, fallback rules apply using only the BI-RADS and density annotations. If all annotation fields are missing or invalid, the corresponding image is not used for model training or validation.

3.4 LLM-Based Evaluation Protocol

To assess the clinical relevance of description produced by vision-language models, we implemented a multi-agent evaluation strategy using accessible LLM APIs. The goal was to simulate expert scoring without human supervision, while ensuring reliable and interpretable judgments.

We employed two primary evaluators:

– **GPT-4o-mini**: a compact version of OpenAI's GPT-4.
– **GPT-4.1-mini**: a lightweight variant of the GPT-4.1 architecture.

Each description was independently evaluated by both models, which were prompted to assign a clinical relevance score from 0 to 4 based on the following rubric:

0: No medical imaging content mentioned
1: Medical content present, but unrelated to mammography
2: Mammography mentioned, but lacking specific clinical detail
3: Specific mammography-related details mentioned
4: Specific mammography content accurately matching the ground truth

If the scores disagreed, a third model, **Gemini 2.5 Flash** [30], was used as a conflict resolver. This model received an adjudication prompt containing the original description, ground truth, both evaluators' scores and justifications, and selected the more appropriate evaluation.

Across the dataset, we observed an **agreement rate of 82%** between the primary evaluators. The remaining 18% were automatically resolved using the conflict resolution mechanism. This pipeline enabled reliable scoring at scale without requiring human annotation. Figure 1 summarizes the evaluation protocol.

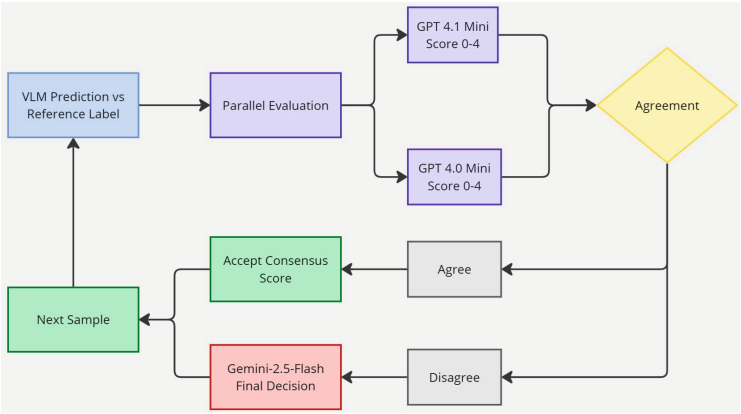


Fig. 1. Evaluation pipeline using two LLMs and a conflict resolver for final adjudication.

4 Results

4.1 Baseline Evaluation

To compare the performance of the selected models, we computed two quantitative metrics based on the LLM-based evaluation pipeline:

- **Average Relevance Score**: The average score assigned by the model across all valid samples (ranging from 0 to 4). It reflects how clinically relevant the generated descriptions are, on average.
- **Normalized Relevance Score**: The average relevance score divided by the maximum possible score (4), resulting in a value between 0 and 1. This normalized form is common in evaluation metrics and facilitates model comparison.

Table 1 presents the initial evaluation results. CLIP+GPT-2 achieved the highest normalized relevance score (0.4687), closely followed by MiniGPT-4 (0.4616). BLIP ranked third (0.3135), demonstrating moderate alignment with clinical content. GIT scored substantially lower (0.0098), indicating suboptimal performance in describing mammographic images.

Table 1. Relevance scores assigned to model-generated descriptions.

Model	Average Relevance Score	Normalized Relevance Score
BLIP	1.2545	0.3135
MiniGPT-4	1.8471	0.4616
GIT	0.0392	0.0098
CLIP+GPT-2	**1.8759**	**0.4687**

4.2 Qualitative Examples

To better illustrate model behavior under identical visual input, we present in Table 2 the predictions of all four evaluated models for the same image (Img 1). This allows a direct comparison of description quality, specificity, and alignment with the reference description shown earlier in Sect. 3.3.

Table 2. Model predictions for the same image (Img 1) from each model.

Image	Model Prediction	Model
Img 1	*a mammogram (confidence: 0.59)*	CLIP+GPT-2
Img 1	*a close up of a breast with a black background*	BLIP
Img 1	*a bright light shining on a dark surface*	GIT
Img 1	*an image of a woman's breasts in a x-ray image*	MiniGPT-4

From this comparison, we observe that CLIP+GPT-2 is the only model to explicitly identify the image as a mammogram, though without detailed clinical descriptors. MiniGPT-4 also produces a structurally coherent and medically plausible description, but lacks specificity and clinical terminology. BLIP provides a generic anatomical reference, while GIT generates a clearly unrelated and hallucinated output. These qualitative insights are consistent with the quantitative scores, reinforcing the overall superiority of CLIP+GPT-2 for fine-tuning.

4.3 Fine-Tuning and Evaluation Protocol

Following the initial evaluation, we selected the CLIP+GPT-2 model, which was identified as the top-performing candidate, for fine-tuning using the full VinDr-Mammo dataset [29]. This choice was motivated by its ability to correctly identify the image modality (i.e., mammograms), even with limited clinical specificity in the baseline.

We implemented a full training and validation pipeline using PNG images and structured descriptions automatically generated from VinDr annotations (see Sect. 3.3). The objective was to improve the model's capacity to produce clinically meaningful descriptions while maintaining computational efficiency, adopting principles of parameter-efficient tuning [31].

To assess performance improvements, we computed BLEU, ROUGE-L, and METEOR metrics on a held-out validation set, in addition to Average Relevance Score and Normalized Relevance Score. These metrics capture different aspects of description quality: n-gram overlap (BLEU), sequence similarity (ROUGE-L), and semantic alignment (METEOR). Baseline scores were derived from the original GPT-2 decoder without visual input. Improvements were monitored across epochs.

The training process includes:

– Resizing all images to 512×512 and normalizing to CLIP's input format (chosen to reduce computational cost);

- Tokenizing descriptions with GPT-2's tokenizer and applying appropriate padding;
- Freezing most CLIP layers and applying LoRA-based adaptation to GPT-2;
- Logging per-epoch metrics, training time, and loss values with a custom logger.

Prior to training, we validated the pipeline with metric checks, forward pass verification, and inference sanity tests. Evaluation was performed at each epoch, with automatic checkpointing. Final results included quantitative scores (see Table 3).

This protocol enables scalable fine-tuning of vision-language models for radiology-focused image-to-text generation, without the need for extensive computing infrastructure, thus promoting broader accessibility in clinical AI applications.

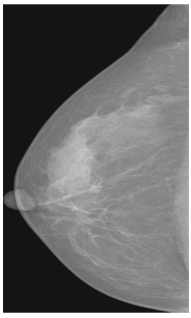


Fig. 2. Example mammogram image used during training.

Figure 2 illustrates a representative mammogram used during training. This craniocaudal (CC) view captures internal breast anatomy with high-resolution contrast, including fibroglandular density and subtle asymmetries. These features represent the type of diagnostic signals the model learns to describe through multimodal learning.

4.4 Post-finetuning Results

To quantitatively assess the impact of the fine-tuning process, the performance of the CLIP+GPT-2 model before and after adaptation using 15,000 training images was compared. Evaluation was conducted on a held-out validation set of 5,000 images. Table 3 summarizes the results.

The results reveal a surprising performance boost across all metrics following fine-tuning. BLEU scores increased by more than 32 times, indicating significantly improved n-gram overlap with the reference descriptions. ROUGE-L, METEOR, Average Relevance Score, and Normalized Relevance Score also exhibited substantial gains, reflecting enhanced structural alignment and semantic fidelity. These improvements validate the effectiveness of our training strategy and demonstrate that open-access, pretrained vision-language models can be

Table 3. Evaluation metrics before and after fine-tuning on 15,000 training images. Metrics computed on 5,000 validation images.

Metric	Baseline (CLIP+GPT-2)	Fine-tuned	Relative Improvement
BLEU	0.0156	0.5277	+3282.6%
ROUGE-L	0.0312	0.6724	+2055.3%
METEOR	0.0489	0.8984	+1737.2%
Average Relevance Score	1.8759	3.4924	+86.2%
Normalized Relevance Score	0.4687	0.8741	+86.5%

successfully adapted for structured clinical description generation when guided by domain-specific supervision.

5 Discussion

Our results demonstrate that open-access vision-language models can be effectively adapted for mammographic description generation, achieving substantial improvements in structured text generation through domain-specific fine-tuning. The performance gains across BLEU, ROUGE-L, METEOR, Average Relevance Score, and Normalized Relevance Score highlight the potential of accessible AI solutions to support radiological workflows. Building on these results, the proposed approach offers a feasible path to adapting open-source models without extensive computational resources, promoting broader clinical adoption.

However, important considerations must be acknowledged. While improved metrics indicate enhanced potential for generating structured mammographic descriptions, these primarily assess linguistic similarity rather than clinical accuracy in capturing specific findings, breast density classifications, or BI-RADS categories. Medical validation would be necessary to ensure diagnostic precision and clinical utility.

Furthermore, our methodology focuses on single mammographic views to maintain computational efficiency and model simplicity. This design choice precludes comprehensive bilateral breast assessment and limits the evaluation of asymmetric findings.

Future work should prioritize clinical accuracy evaluation through expert radiologist assessment to establish medical validation and ensure diagnostic precision. Additionally, extending the framework to incorporate bilateral mammographic analysis would enable comprehensive breast evaluation and asymmetry detection, enhancing clinical utility.

6 Conclusion

This work demonstrates that open-access vision-language models can be successfully adapted for mammographic image description, benefiting significantly

from structured domain supervision. Our evaluation identified CLIP+GPT-2 as the most suitable architecture for clinical adaptation, with fine-tuning yielding substantial gains across all evaluation metrics.

The key contributions include: (1) a comprehensive evaluation framework using multi-agent LLM assessment; (2) demonstration that accessible AI models can achieve clinically relevant performance; (3) a practical fine-tuning methodology that significantly improves description quality while maintaining deployment feasibility; and (4) evidence that modular vision-language architectures may offer advantages in medical imaging applications.

Acknowledgments. This work has been partially funded by the project *Research and Development of Algorithms for Construction of Digital Human Technological Components*, supported by the Advanced Knowledge Center in Immersive Technologies (AKCIT), with financial resources from the PPI IoT/Manufatura 4.0 / PPI HardwareBR of the MCTI, grant number 057/2023, signed with EMBRAPII.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Harvey, H., Glocker, B.: A standardised approach for preparing imaging data for machine learning tasks in radiology. In: Ranschaert, E.R., Morozov, S., Algra, P.R. (eds) Artificial Intelligence in Medical Imaging: Opportunities, Applications and Risks, pp. 61–72. Springer, Cham (2019). https://doi.org/10.1007/978-3-319-94878-2_6
2. Siegel, R.L., Miller, K.D., Wagle, N.S., Jemal, A.: Cancer statistics. CA Cancer J. Clin. **73**(1), 17–48 (2023). https://doi.org/10.3322/caac.21763
3. Lehman, C.D., et al.: National performance benchmarks for modern screening digital mammography: update from the breast cancer surveillance consortium. Radiology **283**(1), 49–58 (2017). https://doi.org/10.1148/radiol.2016161174
4. Branco, P.E.S.C., et al.: Artificial intelligence in mammography: a systematic review of the external validation. Rev Bras Ginecol Obstet **46**, e-rbgo71 (2024). https://doi.org/10.61622/rbgo/2024rbgo71
5. Li, J., Li, D., Xiong, C., Hoi, S.: BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. arXiv preprint arXiv:2201.12086 (2022)
6. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., et al.: Learning transferable visual models from natural language supervision. arXiv preprint arXiv:2103.00020 (2021)
7. Rajpurkar, P., Chen, E., Banerjee, O., Topol, E.J.: Ai in health and medicine. Nat. Med. **28**(1), 31–38 (2022). https://doi.org/10.1038/s41591-021-01614-0
8. OpenAI: GPT-4V(ision) system card. Technical report (2023)
9. Bommasani, R., Hudson, D.A., Adeli, E., et al.: On the opportunities and risks of foundation models. arXiv preprint, arXiv:2108.07258 (2022)
10. Xu, K., Ba, J., Kiros, R., et al.: Show, attend and tell: Neural image caption generation with visual attention. arXiv preprint arXiv:1502.03044 (2015)

11. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: CVPR, pp. 3156–3164 (2015)
12. Jing, B., Xie, P., Xing, E.: On the automatic generation of medical imaging reports. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 2577–2586. Association for Computational Linguistics (2018). https://doi.org/10.18653/v1/P18-1240
13. Chen, Z., Song, Y., Chang, T.H., Wan, X.: Generating radiology reports via memory-driven transformer. arXiv preprint arXiv:2010.16056 (2022)
14. You, D., Liu, F., Ge, S., Xie, X., Zhang, J., Wu, X.: Aligntransformer: hierarchical alignment of visual regions and disease tags for medical report generation. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12903, pp. 72–82. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87199-4_7
15. Miura, Y., Zhang, Y., Tsai, E., Langlotz, C., Jurafsky, D.: Improving factual completeness and consistency of image-to-text radiology report generation. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 5288–5304. Association for Computational Linguistics (2021). https://aclanthology.org/2021.naacl-main.416
16. Kim, H.E., Kim, H.H., Han, B.K., Kim, K.H., Han, K., Nam, H., et al.: Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study. Lancet Digit Health **2**(3), e138–e148 (2020). https://doi.org/10.1016/S2589-7500(20)30003-0
17. Ribli, D., Horváth, A., Unger, Z., et al.: Detecting and classifying lesions in mammograms with deep learning. Sci. Rep. **8**, 4165 (2018). https://doi.org/10.1038/s41598-018-22437-z
18. Chougrad, H., Zouaki, H., Alheyane, O.: Deep convolutional neural networks for breast cancer screening. Comput. Methods Programs Biomed. **157**, 19–30 (2018). https://doi.org/10.1016/j.cmpb.2018.01.011
19. Wang, J., Yang, X., Cai, H., et al.: Discrimination of breast cancer with microcalcifications on mammography by deep learning. Sci. Rep. **6**, 27327 (2016). https://doi.org/10.1038/srep27327
20. Wu, N., Phang, J., Park, J., et al.: Deep neural networks improve radiologists' performance in breast cancer screening. IEEE Trans. Med. Imaging **39**(4), 1184–1194 (2020). https://doi.org/10.1109/TMI.2019.2945514
21. Mohamed, A.A., Berg, W.A., Peng, H., et al.: A deep learning method for classifying mammographic breast density categories. Med. Phys. **45**(1), 314–321 (2018). https://doi.org/10.1002/mp.12683
22. Moor, M., Banerjee, O., Abad, Z.S.H., et al.: Foundation models for generalist medical artificial intelligence. Nature **616**, 259–265 (2023). https://doi.org/10.1038/s41586-023-05881-4
23. Tu, T., Azizi, S., Driess, D., et al.: Towards Generalist Biomedical AI. NEJM AI **1**(3), (2024). https://doi.org/10.1056/AIoa2300138
24. Huang, S.C., Shen, L., Lungren, M.P., Yeung, S.: GLoRIA: a Multimodal Global-Local Representation Learning Framework for Label-efficient Medical Image Recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 3922–3931 (2021). https://doi.org/10.1109/ICCV48922.2021.00391
25. Driess, D., Xia, F., Sajjadi, M.S., et al.: PaLM-E: an embodied multimodal language model. In: ICML, pp. 8469–8488. PMLR (2023)

26. Dai, W., Li, J., Li, D., et al.: InstructBLIP: towards General-Purpose Vision-Language Models with Instruction Tuning. In: Proceedings of the 37th International Conference on Neural Information Processing Systems (NeurIPS), pp. 49250–49267. PMLR (2023)
27. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: MiniGPT-4: enhancing vision-language understanding with advanced large language models. In: Proceedings of the Twelfth International Conference on Learning Representations (ICLR) (2024). https://openreview.net/forum?id=1tZbq88f27
28. Wang, J., Yang, Z., Hu, X., et al.: GIT: a Generative Image-to-Text Transformer for Vision and Language. Trans. Mach. Learn. Res. (2022). https://openreview.net/forum?id=b4tMhpN0JC
29. Pham, H.H., Nguyen Trung, H., and Nguyen, H.Q.: VinDr-Mammo: a large-scale benchmark dataset for computer-aided detection and diagnosis in full-field digital mammography. PhysioNet (2022). Version 1.0.0. https://doi.org/10.13026/br2v-7517
30. Gemini Team, Anil, R., Borgeaud, S., et al.: Gemini: A Family of Highly Capable Multimodal Models. arXiv preprint arXiv:2312.11805 (2025)
31. Hu, E.J., et al.: LoRA: low-Rank Adaptation of Large Language Models. In: International Conference on Learning Representations (2022). https://openreview.net/forum?id=nZeVKeeFYf9
32. Chen, I.Y., Pierson, E., Rose, S., et al.: Foundation models for healthcare: opportunities, risks, and strategies forward. Nat. Med. **28**(5), 890–896 (2022)



Can We Teach AI to Understand Breast Tumour Behaviour? Our MAMA-MIA Challenge Journey

Hadeel Awwad[1(✉)], Joan C. Vilanova[2], and Robert Martí[1]

[1] Computer Vision and Robotics Institute, University of Girona, 17004 Girona, Spain
hadeel.awwad@udg.edu

[2] Department of Radiology, Clínica Girona, Institute of Diagnostic Imaging (IDI), 17005 Girona, Spain

Abstract. Breast cancer is a heterogeneous disease requiring accurate imaging interpretation for diagnosis and treatment planning. Dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI) plays a critical role in this setting by capturing tumour vascular dynamics. In this work, we present our contribution to the MAMA-MIA Challenge, which targets primary tumour segmentation (Task 1) and treatment response prediction from pre-treatment DCE-MRI (Task 2). For Task 1, we adopt the nnU-Net framework, combined with a tailored preprocessing pipeline designed to enhance generalisation and fairness across heterogeneous imaging centres and scanner vendors. Our pipeline focuses on extracting meaningful dynamic information from three representative DCE-MRI phases, used as multi-channel input to guide the model in learning temporal enhancement behaviour around the tumour. For Task 2, we extract temporal dynamics features from all phases and combine them with radiomics-based shape descriptors, derived from the predicted tumour mask, to train an XGBoost classifier for (pathological complete response (pCR)) prediction. Our work aligns with the broader goals of developing robust, generalisable, and equitable AI tools for breast cancer imaging using real-world, multi-centre MRI data.

Keywords: Breast Imaging · Segmentation · DCE-MRI · Treatment response prediction

1 Introduction

Breast cancer remains the most frequently diagnosed cancer and a leading cause of cancer mortality in women worldwide [14]. Neoadjuvant chemotherapy (NAC) is commonly administered before surgery for patients with locally advanced tumours or aggressive subtypes, aiming to downstaging the tumours and improve surgical outcomes. Treatment response is typically assessed by (pCR), defined as the absence of invasive disease in the breast and lymph nodes. Achieving pCR is associated with significantly improved survival, yet only ($\approx 7-38\%$)

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
T. Zhang et al. (Eds.): Deep-Breath 2025, LNCS 16142, pp. 248–257, 2026.
https://doi.org/10.1007/978-3-032-05559-0_25

of patients attain this outcome under standard regimens [14]. Early identification of likely responders is thus crucial for enabling therapy adaptation and minimising unnecessary toxicity of ineffective NAC [14]. Imaging plays a central role in both baseline tumour characterisation and response monitoring. Dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI) is particularly effective for lesion detection and staging, and is considered the most accurate modality for evaluating NAC response by measuring changes in tumour volume [4,14]. Changes in kinetic parameters, such as the contrast transfer constant K_{trans}, have shown predictive value mid-treatment [14]; however, the ability to predict pCR from pre-treatment imaging alone remains an open challenge. Few studies have investigated this task using pre-NAC DCE-MRI, often with small cohorts and modest performance (area-under-curve (AUC) 0.56−0.66) [14], highlighting the need for more advanced, data-driven approaches. Precise tumour segmentation is a critical step for quantitative breast MRI analysis, enabling extraction of metrics such as volume and shape, which are essential for radiomics and therapy monitoring [4]. Manual segmentation is labour-intensive, restricting many studies to small datasets. Publicly available annotated datasets are rare–for instance, I-SPY1 contains only 163 DCE-MRI cases with segmentations [4]. Automated segmentation has advanced with deep learning (DL), particularly Convolutional neural networks (CNNs). Early models trained on limited datasets ($\sim 50 - 250$ MRIs) achieved Dice scores (DSC) around $0.60 - 0.75$ [6]. Recent large-scale studies have demonstrated radiologist-level performance, with 3D U-Nets achieving a median DSC of ~ 0.77–0.87, strong tumour volume correlation ($R \approx 0.95$), and good discrimination of treatment response (AUC ~ 0.76) [6,9]. These findings confirm the feasibility of reliable pre-NAC segmentation and its importance for downstream predictive modelling.

Traditional prediction of pCR has used radiomics-handcrafted features derived from pre-therapy DCE-MRI and used machine learning classifiers. Reported AUCs range from 0.57 to 0.79 [14], though performance is often limited by feature design and dataset size. Recent DL-based methods that learn features directly from imaging achieve similar AUC ranges ($\sim 0.55 - 0.80$) with one study showing baseline MRI outperforming post-NAC MRI (AUC ~ 0.80 vs ~ 0.77) [14]. This suggests pre-treatment imaging may encode key biological traits relevant to treatment response, though current models lack sufficient accuracy for clinical deployment.

This work proposes a unified framework for tumour segmentation and pCR prediction using pre-treatment DCE-MRI. We propose a two-stage pipeline comprising: automated primary tumour segmentation via nnU-Net and treatment response prediction using radiomics-based shape features and dynamic descriptors to capture tumour enhancement behaviour. Our results are based on the MAMA-MIA Challenge validation phase, evaluated on a private cohort of 58 patients from three European centres. This work contributes toward developing reproducible, generalisable, and fair AI tools for breast cancer management.

2 Materials

We conducted our experiments on the MAMA-MIA dataset [4], a large-scale, multi-institutional benchmark comprising 1,506 pre-treatment DCE-MRI cases collected from over 20 U.S. clinical centres. The dataset aggregates scans from four TCIA repositories: DUKE [16], I-SPY1 [12], I-SPY2 [10], and the Single-Site Breast NACT dataset [11]. Each case includes a multi-phase, T1-weighted DCE-MRI volume, expert-annotated 3D tumour segmentation, and a binary pCR label. Additionally, 49 harmonized clinical and acquisition-related metadata variables are provided, including age, menopausal status, tumour subtype, and scanner details. The challenge organisers defined a fixed public split of 1,200 training and 306 internal testing cases [3], along with a hidden validation set of 58 European cases for on-platform evaluation. An additional held-out test cohort comprising 516 cases is reserved for final challenge evaluation. Ground truth for both the validation and final test cohorts remains undisclosed; only quantitative performance metrics are returned post-submission. Bounding-box annotations of the affected breast were available for all subsets. During model development, two instances with mislocalized segmentation annotations were identified. These were manually corrected and subsequently validated by an experienced radiologist before inclusion in the final training set.

3 Methods

Our two-stage pipeline (Fig. 1) leverages pre-treatment DCE-MRI for lesion segmentation and outcome prediction. The first stage (Fig. 1a) employs a 3D nnU-Net for primary tumour segmentation. Subsequently, the second stage (Fig. 1b) utilises the predicted tumour masks to extract quantitative imaging features for (pCR) prediction. All stages are fully automated, operate exclusively on pre-treatment imaging data, and exclude clinical variables as per challenge rules.

3.1 Primary Tumour Segmentation

Preprocessing. To harmonise multi-centre DCE-MRI data while preserving tumour enhancement dynamics. First, we selected three representative temporal phases: pre-contrast, first post-contrast, and second post-contrast. This selection ensured phase consistency across all splits and aligned with prior studies that demonstrated the importance of early post-contrast kinetics for lesion characterisation [2]. To spatially focus the input, we utilised the provided breast bounding-box coordinates to crop each phase to the affected breast. This step reduces computational burden and GPU memory usage, improves signal-to-noise ratio (SNR), and minimises false positives originating from the contralateral breast [15]. Following cropping, we applied percentile-based intensity clipping (0.1–99.9%) to suppress scanner-dependent outliers and reduce the influence of extreme noise or artefacts. This has been shown to stabilise intensity distributions and facilitate normalisation across heterogeneous sources [13]. We then

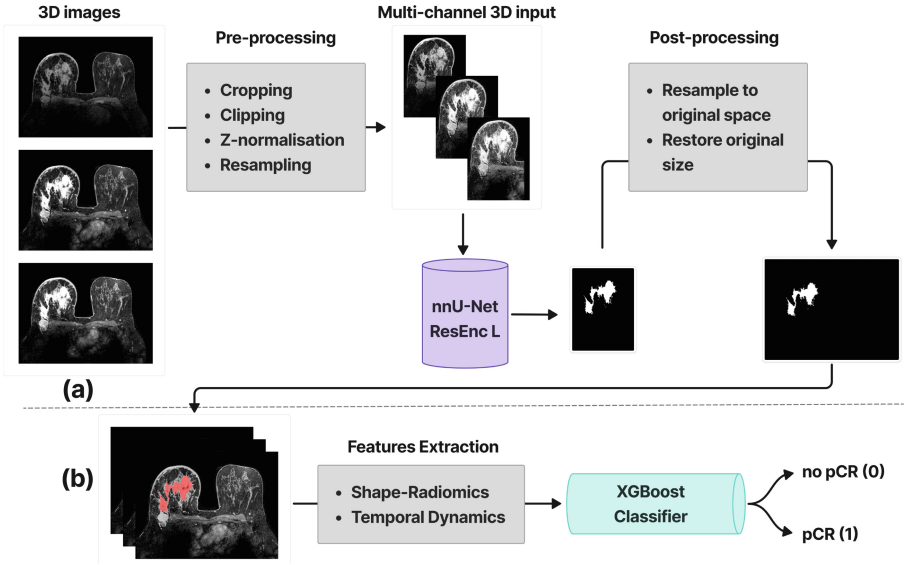


Fig. 1. Overview of the proposed framework. (a) Primary tumour segmentation using nnU-Net ResEnc L. (b) Treatment response prediction using features extracted from the predicted tumour region.

performed Z-score normalisation using the pre-contrast phase's mean and standard deviation as a reference. This phase serves as a robust baseline unaffected by contrast uptake and enables standardised intensity scaling across patients. Lastly, we resampled all images to an isotropic voxel spacing of 1 mm using B-spline interpolation. This resampling step ensures spatial consistency for 3D convolutional operations. The same preprocessing pipeline was applied at inference time to maintain consistency between the training and evaluation phases.

Segmentation Model Architecture. We adopted the nnU-Net framework [7] for primary tumour segmentation, using the ResEnc L configuration [8], which preserves the full-resolution 3D U-Net architecture while enhancing representational capacity via residual blocks and increased depth. In contrast to the default nnU-Net planner, this configuration uses a redesigned planning strategy that adapts key hyperparameters (e.g., target spacing, patch size, batch size) to each dataset's fingerprint, thereby improving the model's ability to generalise across imaging conditions. We use three DCE-MRI phases pre-contrast, first post-contrast, and second post-contrast as multi-channel 3D input, consistent with the preprocessing described earlier. No manual modifications were made to the architecture or training settings beyond selecting the ResEnc L preset. The model was trained for 1000 epochs on the official train/test split provided by the

MAMA-MIA Challenge organisers, ensuring a consistent evaluation protocol on a hold-out validation set.

Postprocessing. To restore predicted segmentations to the original image's coordinate system, a postprocessing step was implemented. This involved resampling the cropped and isotropically-resampled predictions back to their native voxel spacing and repositioning them into the full-size volume using metadata from the initial cropping, facilitated by the SimpleITK library [17]. This ensures spatial consistency for accurate analysis, comparison with ground-truth, and clinical application..

3.2 Treatment Response Prediction

Feature Extraction. To predict pCR, we engineered a comprehensive feature set from DCE-MRI scans, focusing on morphological and haemodynamic properties given their established biological links to tumour aggressiveness and treatment response. Morphological features, specifically shape-based radiomics, were extracted directly from binary tumour segmentation masks using PyRadiomics [5], providing geometry-specific insights. Temporal dynamics features (Baseline mean intensity, wash-in rate, wash-out slope, area under the curve (AUC), intensity range, and voxel mean over time) were derived from mean tumour intensity enhancement curves, calculated from the tumour region across the entire DCE-MRI phases, to capture perfusion characteristics. All these extracted features were included to provide the subsequent classifier with a comprehensive set of descriptors, allowing the learning algorithm to determine the most predictive patterns without premature feature reduction. We deliberately excluded other radiomics categories. This choice prioritised shape features for their superior robustness across various MRI scanners and acquisition parameters, and recognised that our custom temporal features more comprehensively capture dynamic intensity information. This approach creates a more streamlined and interpretable model by clearly distinguishing static morphological properties from dynamic haemodynamic function. For cases with empty segmentation masks, all feature values were set to zero and a (segmentation_failed) binary flag was introduced. The classifier was configured to default to a non-pCR outcome with a 0.5 probability for such cases. This strategy was employed to maintain robustness in downstream prediction and, importantly, to avoid introducing noise from potentially unreliable or absent segmentations. The selected features and their relevance to pCR prediction are detailed in Appendix 1.

Model for pCR Prediction. These extracted imaging features were then used to classify treatment response using an XGBoost classifier [1] with a binary logistic objective. XGBoost was chosen over simpler models due to its inherent ability to model complex non-linear relationships and interactions among diverse imaging features, capturing intricate patterns that linear approaches

often miss. Despite its complexity and the limited number of features, overfitting was actively mitigated through specific configurations and robust evaluation strategies. The model was trained with 500 estimators and a maximum tree depth of 5, a crucial regularization parameter designed to prevent individual trees from becoming overly complex. To enhance stability in the presence of class imbalance, we set the maximum delta step to 1, which constrains leaf weight updates and promotes conservative convergence. Robust evaluation was ensured via five-fold stratified cross-validation, preserving the original distribution of treatment response labels and mitigating sampling bias. Furthermore, cases with missing or unknown pCR labels were assigned to the non-pCR class, in line with the challenge scoring formula [3]. This classifier effectively integrates geometric descriptors and dynamic enhancement patterns for pCR prediction from baseline imaging alone.

3.3 Evaluation Metrics

Evaluation for both tasks followed the official MAMA-MIA Challenge rules [3], combining performance and fairness metrics. Segmentation performance was assessed by a composite score averaging DSC (volumetric overlap) and Normalised 95th Percentile Hausdorff Distance (NormHD) (boundary agreement). For treatment response prediction, Balanced Accuracy (average of true positive and true negative rates) was used to account for class imbalance. Fairness, measured across clinically relevant subgroups (age, menopausal status, breast density), reflected consistency in DSC and NormHD for segmentation, and quantified differences in true positive rates (Equalised Odds Disparity) for prediction. Final scores were computed as the average of performance and fairness scores, promoting equitable outcomes.

Implementation Details. The nnU-Net ResEnc L model was trained on a single NVIDIA A40 GPU with 46 GB of memory. During training on the provided single-fold split, peak GPU memory usage was approximately 27,178 MiB (∼26.5 GiB), as measured using nvidia-smi. For the treatment response prediction task based on XGBoost, model training was performed using CPU resources, as no GPU acceleration was required.

4 Results and Discussion

Evaluation was conducted on two subsets: an internal hold-out test set (306 cases from the challenge's public 1,200/306 split) and a hidden validation set (58 cases from external European sites). The reported total execution time (15.39 min) corresponds solely to generating predictions on this hidden validation set. Results for primary tumour segmentation and treatment response prediction are summarised in Tables 1 and 2, respectively.

In the segmentation task (Table 1), our model achieved a DSC of 0.81 ± 0.16 and a NormHD of 0.06 ± 0.10 on the internal test set (with performance and

fairness scores of 0.88 each). For robust generalisation assessment, we evaluated our model on the hidden validation set, which serves as the independent external evaluation set for the MAMA-MIA challenge. On this set, our model obtained a DSC of 0.71 ± 0.23, a NormHD of 0.10 ± 0.16, and a final score of 0.84. Crucially, this same external dataset was used to evaluate the MAMA-MIA challenge's nnU-Net baseline model (Table 1, row "Baseline (External)"), which achieved a DSC of 0.65, a NormHD of 0.3, and a final score of 0.74. Our model's performance on this shared external set significantly surpasses the MAMA-MIA baseline across all metrics, thereby demonstrating superior generalization and robustness to unseen data. Furthermore, our method achieved the highest DSC and lowest NormHD on the public MAMA-MIA leaderboard, aligning with state-of-the-art 3D U-Nets (median DSC of ∼0.77–0.87, tumour volume correlations around R≈0.95 [4]).

Table 1. Segmentation results across evaluation sets. *Standard deviation is not available.

Evaluation Set	DSC	NormHD	Performance Score	Fairness Score	Final Score
Internal Test	0.81 ± 0.16	0.06 ± 0.10	0.88	0.88	0.88
Hidden Validation	0.71 ± 0.23	0.10 ± 0.16	0.80	0.87	0.84
Baseline (External)*	0.65	0.3	0.68	0.8	0.74

For treatment response prediction (Table 2), our model's performance on baseline DCE-MRI features alone reflected the inherent challenge of predicting pCR. Achieving balanced accuracies near chance (0.51 ± 0.02 internal; 0.46 hidden validation), our results align with the generally limited performance observed across the MAMA-MIA leaderboard, underscoring the task's difficulty. Specifically, on the internal test split, the model obtained a balanced accuracy of 0.51 ± 0.02 and a fairness score of 0.55 ± 0.10 (final score 0.53 ± 0.05). On the hidden validation set, representing an independent external evaluation, our model achieved a balanced accuracy of 0.46, a fairness score of 0.37, and a final score of 0.41. Notably, the MAMA-MIA challenge baseline prediction model, evaluated on this same external dataset (Table 2, row "Baseline (External)"), reported a performance score of 0.48, a fairness score of 0.08, and a final score of 0.28. While both our model and the MAMA-MIA baseline demonstrate limited overall predictive power for pCR from baseline imaging, our model's performance on the hidden validation set (final score 0.41) notably surpasses the baseline (final score 0.28) on this shared external evaluation, particularly in terms of fairness. It needs to be noted that the prediction model was trained using features extracted from pre-treatment DCE-MRI only, without incorporating clinical metadata.

Qualitative segmentation examples illustrating model performance are presented in Fig. 2. Example (a) shows high overlap between predicted and ground truth tumour regions, demonstrating strong agreement. In contrast, (b) shows a challenging case of under-segmentation with reduced spatial agreement. These examples highlight typical variations in model behaviour across the dataset.

Table 2. Treatment response classification results across evaluation sets.* Standard deviation is not available.

Evaluation Set	Balanced Accuracy	Fairness Score	Final Score
Internal Test	0.51 ± 0.02	0.55 ± 0.10	0.53 ± 0.05
Hidden Validation*	0.46	0.37	0.41
Baseline (External)*	0.48	0.08	0.28

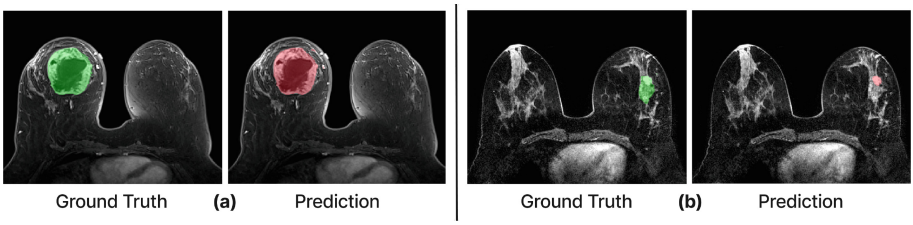


Fig. 2. Qualitative segmentation results on pre-treatment DCE-MRI. (a) Accurate prediction closely matching the ground truth (DSC = 0.96). (b) Under-segmentation example (DSC = 0.53). Ground truth in green, prediction in red. (Color figure online)

5 Conclusion

We presented a two-stage pipeline for primary tumour segmentation and pCR prediction using pre-treatment DCE-MRI in the MAMA-MIA challenge. Our segmentation model, built on the nnU-Net with Residual Encoder, ranked second on the public leaderboard and demonstrated strong performance and fairness across internal and external sites. While the pCR prediction, based solely on pre-treatment imaging features, yielded limited predictive power with near-chance accuracy, this outcome highlights the significant challenge of predicting pCR from baseline MRI alone. Nevertheless, our work demonstrates the feasibility of investigating baseline MRI for early response assessment, underscoring the critical need for more robust, fair, and generalisable prediction models in future research, possibly integrating advanced feature representations or using other classifier models.

Acknowledgement. This research was funded by the project VICTORIA, "PID2021-123390OB-C21" from the Ministerio de Ciencia e Innovación y Universidades of Spain, and the PhD grant IFUdG2024 from the University of Girona.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

Appendix 1

Table 3 provides a comprehensive summary of the engineered features used for pCR prediction, detailing their categories, names, and biological relevance.

Table 3. Summary of Selected Features and Their Relevance to pCR Prediction.

Feature Category	Feature Names	Relevance to pCR Prediction
Morphological	Elongation, Flatness, Least Axis Length, Major Axis Length, Maximum 2D Diameter Column, Maximum 2D Diameter Row, Maximum 2D Diameter Slice, Maximum 3D Diameter, Mesh Volume, Minor Axis Length, Sphericity, Surface Area, Surface Volume Ratio, Voxel Volume	Quantify tumour size, shape complexity, and compactness. Irregular shapes and larger tumour volumes typically indicate more aggressive tumours, suggesting a lower likelihood of pCR.
Temporal Dynamic	Baseline mean intensity, Intensity Peak, Area Under the intensity Curve, intensity range, Wash in, Wash out, Voxel intensity mean over time	Characterise tumour perfusion and vascularity. Features like rapid wash-in and high peak signal suggest increased vascularity, often associated with aggressive cancers. Wash-out dynamics and overall AUC reflect contrast kinetics, providing insights into treatment response.

References

1. Chen, T., Guestrin, C.: Xgboost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794 (2016). https://doi.org/10.1145/2939672.2939785
2. Eskreis-Winkler, S., et al.: Breast mri background parenchymal enhancement categorization using deep learning: outperforming the radiologist. J. Magn. Reson. Imaging **56**(4), 1068–1076 (2022). https://doi.org/10.1002/jmri.28111. epub 2022 Feb 15 Oct
3. Garrucho, L., et al.: Advancing generalizability and fairness in breast mri tumour segmentation and treatment response prediction (mama-mia), March 2025. https://doi.org/10.5281/zenodo.15052678

4. Garrucho, L., et al.: A large-scale multicenter breast cancer dce-mri benchmark dataset with expert segmentations. Sci. Data **12**(1), 453 (2025). https://doi.org/10.1038/s41597-025-04707-4
5. Griethuysen, J.J.M., et al.: Computational radiomics system to decode the radiographic phenotype. Can. Res. **77**(21), e104–e107 (2017). https://doi.org/10.1158/0008-5472.CAN-17-0339. Nov
6. Hirsch, L., et al.: Radiologist-level performance by using deep learning for segmentation of breast cancers on mri scans. Radiol. Artif. Intell. **4**(1), e200231 (2022). https://doi.org/10.1148/ryai.200231
7. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat. Methods **18**(2), 203–211 (2021)
8. Isensee, F., et al.: nnU-Net revisited: a call for rigorous validation in 3d medical image segmentation (2024)
9. Janse, M.H.A., et al.: Deep learning-based segmentation of locally advanced breast cancer on mri in relation to residual cancer burden: a multi-institutional cohort study. J. Magn. Reson. Imaging **58**(6), 1739–1749 (2023). https://doi.org/10.1002/jmri.28679
10. Li, W., et al.: I-SPY 2 breast dynamic contrast enhanced MRI trial (ISPY2) (Version 1) (2022). https://doi.org/10.7937/TCIA.D8Z0-9T85, dataset published on The Cancer Imaging Archive (TCIA)
11. Newitt, D., Hylton, N.: Single site breast dce-mri data and segmentations from patients undergoing neoadjuvant chemotherapy (version 3) (2016). https://doi.org/10.7937/K9/TCIA.2016.QHsyhJKy
12. Newitt, D., Hylton, N., the I-SPY 1 Network, Team, A..T.: Multi-center breast dce-mri data and segmentations from patients in the i-spy 1/acrin 6657 trials (2016). https://doi.org/10.7937/K9/TCIA.2016.HdHpgJLK
13. Nyúl, L.G., Udupa, J.K.: On standardizing the MR image intensity scale. Magn. Reson. Med. **42**(6), 1072–1081 (1999). https://doi.org/10.1002/(sici)1522-2594(199912)42:6⟨1072::aid-mrm11⟩3.0.co;2-m
14. Peng, Y., et al.: Pretreatment dce-mri-based deep learning outperforms radiomics analysis in predicting pathologic complete response to neoadjuvant chemotherapy in breast cancer. Front. Oncol. **12**, 846775 (2022). https://doi.org/10.3389/fonc.2022.846775
15. Ribli, D., Horváth, A., Unger, Z., Pollner, P., Csabai, I.: Detecting and classifying lesions in mammograms with deep learning. Sci. Rep. **8**, 4165 (2018). https://doi.org/10.1038/s41598-018-22437-z
16. Saha, A., et al.: Dynamic contrast-enhanced magnetic resonance images of breast cancer patients with tumor locations (2021). https://doi.org/10.7937/TCIA.e3sv-re93
17. Yaniv, Z., Lowekamp, B.C., Johnson, H.J., Beare, R.: Simpleitk image-analysis notebooks: a collaborative environment for education and reproducible research. J. Digit. Imaging **31**, 290–303 (2018). https://doi.org/10.1007/s10278-017-0037-8

MM-DETR: Emulating the Diagnostic Clinical Workflow in Multi-view Multi-modal Mammography Mass Detection

Karim Elbarbary[1(✉)], Adarsh Bhandary Panambur[1], Sheethal Bhat[1,2], Siming Bayer[1,2], and Andreas Maier[1]

[1] Friedrich-Alexander-Universität, 91058 Erlangen, Germany
karim.khalifa.elbarbary@fau.de
[2] Siemens Healthineers, 91052 Erlangen, Germany

Abstract. Accurate mass detection in dense breast tissue remains a critical challenge in mammographic screening, as overlapping structures often obscure subtle lesions. We observe that a multi-modal vision-language detection framework with strong pretraining over natural images demonstrates remarkable state-of-the-art (SOTA) performance in lesion detection tasks via transfer learning. However, current SOTA does not take advantage of the entire context available to radiologists. Typical mammogram examinations contain two views of each breast for a given patient, and certain lesions are visible in one view, while hidden by dense breast tissue in another. We propose a novel Multi-view Multi-modal DETR object detection framework that trains a network to detect lesions after considering both views, thus emulating the workflow of a radiologist. Specifically, our method incorporates a bidirectional cross-attention fusion module to integrate relevant information from cranio-caudal (CC) and mediolateral oblique (MLO) views simultaneously, reinforcing lesion-specific signals and aiding detection of masses that may be obscured by dense tissue. We evaluate our proposal on the public VinDR-Mammo dataset, achieving significant improvements in mass detection with reduced false negatives. Our method reaches a mass detection mAP of 0.654, outperforming Mammo-CLIP (0.580) by an absolute margin of 12.8%. It also reduces the false negative rate in DENSITY C cases by 5.9% compared to the single-view baseline, highlighting the clinical value of our method. The code is available at https://github.com/MMDETR/MM-DETR.

Keywords: Computer Aided Diagnostics · Multi-modal DETR · Lesion detection · Mammography

1 Introduction

Breast cancer is the most commonly diagnosed cancer and the leading cause of cancer-related death among women, accounting for over 2.3 million new cases

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
T. Zhang et al. (Eds.): Deep-Breath 2025, LNCS 16142, pp. 258–267, 2026.
https://doi.org/10.1007/978-3-032-05559-0_26

and 666,000 deaths worldwide in 2022 [1]. Early detection is crucial for improving clinical outcomes, as the five-year survival rate exceeds 99% when the disease is diagnosed at an early stage [2]. To enable early detection at scale, mammography, capturing CC and MLO views with low-dose X-rays, has become the clinical gold standard for breast cancer screening [3]. *Masses* are among the most critical radiological indicators of malignancy in mammograms, appearing as localized, space-occupying lesions that require accurate detection and assessment [4]. However, in populations with a high prevalence of dense breast tissue, such as those in Asia, there is an increased risk of missed detections of mass lesions, as dense tissue can obscure the lesions, leading to a higher false negative rate in mammography [5]. Moreover, it is also recognized as an independent risk factor for breast cancer development [6].

Early computer-aided detection (CAD) systems were developed for identifying abnormalities in mammography, but their reliance on handcrafted features led to a shift toward more advanced solutions [7]. With advances in deep learning, CNN-based methods have become the dominant approach, often framing mammogram analysis as an image-level classification task [8,9]. However, such models lacked precise localization capabilities. Subsequent works adopted object detection models like Faster R-CNN [10] and YOLOv8 [11] to localize lesions via bounding boxes [12,13], offering improved spatial precision. Yet, the need to downscale high-resolution mammograms often leads to missed small or subtle lesions [14]. Recent studies have proposed multi-view models that jointly process both CC and MLO views, leveraging their complementary perspectives to improve lesion detection [15,16]. These models have demonstrated improved performance in detecting subtle findings that may be missed when relying on a single projection. However, their effectiveness in dense breast tissue remains largely unexplored. Additionally, these works are proprietary. More recently, vision-language models (VLMs), which align visual and textual representations, have achieved state-of-the-art performance in various open-vocabulary object detection tasks [17,18]. Their strong generalization ability, enabled by large-scale pretraining, makes them highly transferable across domains. In medical imaging, VLMs have shown promising results in chest X-ray domain [19,20], yet their use in mammography remains limited. To date, only Mammo-CLIP [21] has explored using a pretrained VLM, which employs image-text alignment, and then leveraging the vision backbone from this pretraining to build a dedicated detector for object detection in mammograms.

Building on the strong single-view detection capabilities of a multi-modal VL DETR framework called Grounding DINO [18], we introduce Multi-view Multi-modal DETR (MM-DETR) that includes both views of the breast for lesion detection.

MM-DETR incorporates a bidirectional cross-attention fusion module that processes both CC and MLO views of a breast image simultaneously. We designate a supporting MLO or CC view for its corresponding primary view counterpart, thereby improving the network's ability to capture inter-view dependencies and contextual cues relevant to the prediction task. This is especially relevant considering dense breast tissues where radiologists need to evaluate both views to detect lesions or masses.

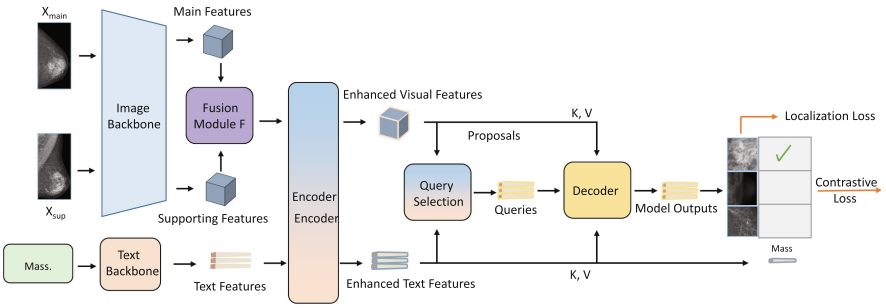


Fig. 1. Overview of MM-DETR. The main view X_{main} and supporting view X_{sup} are processed through a shared Swin Transformer backbone to extract visual features. Bidirectional cross-attention is then applied between the two views to generate a fused representation. This is combined with a text embedding of the fixed prompt "Mass" and passed to the transformer decoder to guide prompt-based detection. The same prompt is used during both training and inference. For negative cases (i.e., images without masses), the model does not produce any bounding boxes in response to the prompt.

We evaluate MM-DETR on a public VinDR-Mammo [22] dataset, demonstrating significant improvement over the single-view baseline and other SOTA methods. We further analyze the results to investigate the density-wise impact and BI-RADS-level stratification, thereby verifying the importance of context in mass detection in dense breast images.

Lastly, several different approaches to the multi-view fusion strategy are assessed. We investigate fusing the multiple views, such as dot product, concatenation, cross-attention and bidirectional cross-attention to integrate complementary information between mammographic projections. This is presented as an ablation study to evaluate the impact of different fusion strategies on detection performance.

Main Contributions: 1) We introduce a Multi-View Multi-Modal DETR (MM-DETR) framework that leverages additional contextual information, replicating the clinical diagnostic process. MM-DETR is implemented over a strong baseline Grounding DINO [18], achieving significant improvement in mass detection on a public VinDR-Mammo dataset [22]. 2) We conduct a comprehensive evaluation of the impact of multi-view learning on dense breast images, thereby quantifying the density-wise improvements. 3) We evaluate additional variations in multi-view fusion strategies, providing ablation studies that qualify the final design choices.

2 Method

Grounding DINO [18] implements a multi-modal vision-language object detection framework. Powerful pretraining over 1.8 million annotations, along with

automatically generated captions, enables zero-shot open-set vocabulary detection over unseen categories. This remarkable performance is achieved through a cross-attention mechanism between the vision and language embeddings in the DETR [23] transformer encoder and contrastive learning with language-guided detection in the decoder. We find that applying grounding DINO [18] on a mammography dataset yields impressive results for mass detection, outperforming previous SOTA methods.

We propose to extend this method to consider both the views (CC and MLO) captured during a mammogram examination, as illustrated in Fig. 1. This provides additional contextual cues to the network by simultaneously processing the CC and MLO views, thereby leveraging complementary anatomical perspectives. This is accomplished by the introduction of a "Multi-view fusion module $\mathcal{F}$" that implements bidirectional cross-attention between the CC and MLO embeddings while preserving alignment in the multi-modal latent space. A fixed prompt, "Mass", is used as the text input during both training and inference. It is encoded using a pretrained text encoder and fused with the multi-view visual features. The model is trained to localize masses when present and to suppress predictions when absent, despite receiving the same prompt in both cases.

2.1 Multi-view Fusion Module $\mathcal{F}$

This module fuses the information from both views to create a single representative embedding containing the necessary lesion information to be detected. This is accomplished by denoting one of the views as the main view and the other as a supporting view. For example, for a main CC view, the corresponding MLO view is the supporting view, and the image backbone extracts feature embeddings for both views simultaneously.

Let $\mathbf{X}_{\text{main}}, \mathbf{X}_{\text{sup}} \in \mathbb{R}^{n_i \times d}$ represent the sets of patch features from the main and supporting views. To fully leverage inter-view dependencies, we implement bidirectional cross-attention by allowing each view to attend to the other. The fused representation $\mathbf{X}_{\text{fused}} = \mathcal{F}(\mathbf{X}_{\text{main}}, \mathbf{X}_{\text{sup}})$ is computed as follows:

$$\mathbf{X}_{\text{fused}} = \mathbf{Z}_{\text{main}\leftarrow\text{sup}} + \mathbf{Z}_{\text{sup}\leftarrow\text{main}}, \tag{1}$$

where, $\mathbf{Z}_{\text{main}\leftarrow\text{sup}}$ and $\mathbf{Z}_{\text{sup}\leftarrow\text{main}}$ are the cross-attended representations obtained by allowing each view to attend to the other:

$$\mathbf{Z}_{\text{main}\leftarrow\text{sup}} = \text{CrossAttn}(\mathbf{Q} = \mathbf{X}_{\text{main}}, \mathbf{K} = \mathbf{X}_{\text{sup}}, \mathbf{V} = \mathbf{X}_{\text{sup}}), \tag{2}$$

$$\mathbf{Z}_{\text{sup}\leftarrow\text{main}} = \text{CrossAttn}(\mathbf{Q} = \mathbf{X}_{\text{sup}}, \mathbf{K} = \mathbf{X}_{\text{main}}, \mathbf{V} = \mathbf{X}_{\text{main}}). \tag{3}$$

This bidirectional fusion mechanism enables the model to jointly reason across both projections, enhancing its ability to localize lesions that may only be partially visible or ambiguous in one view alone.

In both single-view and multi-view settings, the model relies on a language-guided query selection mechanism [18] that selects the most informative image

tokens to serve as decoder queries. Text and image features are first projected into a shared embedding space. Then, pairwise similarity between each image token and the textual prompt is computed using a dot product. The top n image features with the highest similarity are retained and passed as decoder queries. Formally, this is given by:

$$\mathbf{i}_n = \mathrm{Top}_n \left(\overset{(-1)}{\max} \left(\mathbf{X}_{\text{fused}} \mathbf{X}_t^\top \right) \right), \tag{4}$$

where $\mathbf{X}_{\text{fused}} \in \mathbb{R}^{n_i \times d}$ and $\mathbf{X}_t \in \mathbb{R}^{n_t \times d}$ are the image and text features, respectively. The filtered queries $\mathbf{i}_n$ are then passed to the transformer decoder along with the image features $\mathbf{X}_{\text{fused}}$, where they jointly attend to visual and textual information to produce detection predictions.

The model is trained end-to-end using a weighted sum of classification, box regression, and IoU losses:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{cls}} + \lambda_2 \mathcal{L}_{\text{L1}} + \lambda_3 \mathcal{L}_{\text{GIoU}}, \tag{5}$$

where $\mathcal{L}_{\text{cls}}$ is the focal loss [24] for classification, $\mathcal{L}_{\text{L1}}$ is the L1 loss on bounding box coordinates, and $\mathcal{L}_{\text{GIoU}}$ is the generalized IoU loss [25]. The coefficients λ_1, λ_2 and λ_3 control the relative weighting of each loss component.

3 Experimental Setup

3.1 Dataset Description

We train and evaluate our method on the VinDr-Mammo dataset [22], which comprises 20,000 full-field digital mammography images from 5,000 patients collected in Vietnam[1] due to the prevalence of higher breast density in this population. Each patient is imaged in two standard projections, CC and MLO, for both breasts. The dataset includes bounding box annotations for various lesion types, such as masses, calcifications, and architectural distortions. In this study, we restrict our focus to the "Mass" class, following the data configuration established by Mammo-CLIP [21] for comparability. The breast density distribution is heavily skewed toward dense tissue, with 76.46% of the images labeled as density C and 13.50% as density D, while lower-density categories B and A account for only 9.54% and 0.50% of the dataset, respectively.

3.2 Implementation Details

All experiments were conducted using the MMDetection framework [27] on a single node with 48.6 GB Quadro RTX 8000 GPUs. The original text encoder is replaced with Bio_ClinicalBERT [28], which is pretrained on clinical notes. MM-DETR utilizes a modified COCO-format [29] dataset and a custom pipeline to

[1] Vietnam is an Asian country where a higher prevalence of dense breast tissue has been reported among women [26].

load paired CC and MLO views. The models were trained for 40 epochs using the AdamW optimizer, an adaptation of the Adam optimizer [30], with a base learning rate of 5×10^{-5}. Both backbones were frozen during training. Batch size was set to 8 for single-view and 1 for multi-view models. Object queries were set to 300. The loss function included Focal Loss ($\alpha = 0.25$, $\gamma = 2.0$, $\lambda_1 = 2.0$), L1 loss ($\lambda_2 = 5.0$), and GIoU loss ($\lambda_3 = 2.0$). During inference, images were resized to $(800, 1333)$ with aspect ratio preserved, and predictions below a confidence threshold of 0.3 were discarded. Final mAP@0.5 results were averaged over three runs using the same train/test split from the filtered VinDr-Mammo dataset [22].

Table 1. Comparative mass detection performance (mAp) of vision-language models and vision-based baselines on the VinDr-Mammo dataset [22]. Values indicate mean mAP across three runs with standard deviation, except where unavailable.

Model	Mass (mAP)
YOLOv3 [31]	0.523 ± 0.024
Faster R-CNN [10]	0.556 ± 0.003
RetinaNet [24]	0.390
Mammo-CLIP [21]	0.580
Grounding DINO (Single-View)	0.623 ± 0.009
MM-DETR (Ours)	**0.654 ± 0.002**

4 Results and Discussion

Table 1 presents the detection results for mass lesions on the VinDr-Mammo dataset. The evaluation includes both vision-based and vision-language baselines, as well as our proposed single-view and multi-view approaches. We report the mean average precision (mAP) at an IoU threshold of 0.5, averaged over three to five runs.

The results indicate that Grounding DINO outperformed Mammo-CLIP, achieving a mean average precision (mAP@0.5) of 0.623, which represents a 9.8% improvement. This gain can be attributed to Grounding DINO's architecture, which integrates text and image features jointly across three key stages: feature encoding, query selection, and prediction refinement, enabling more precise localization of lesion-relevant regions. Building on this, MM-DETR further improved performance, achieving an mAP of 0.654 and yielding a 5.0% gain over the single-view Grounding DINO. By incorporating complementary anatomical information from both CC and MLO views through bidirectional cross-attention fusion, MM-DETR effectively consolidates spatial cues across projections, enhancing the model's ability to detect subtle or ambiguous lesions with greater consistency.

Furthermore, we perform a comparative analysis of false negative rates across breast density and BI-RADS categories, highlighting the effectiveness of MM-DETR in reducing detection errors. As shown in Fig. 2, MM-DETR lowered the

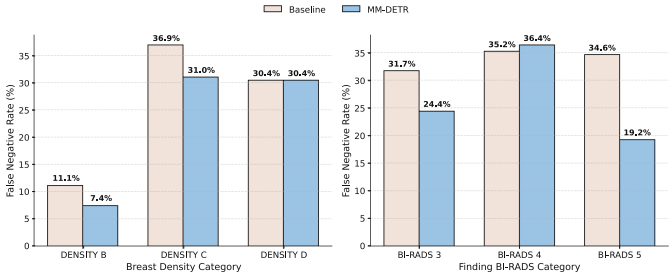


Fig. 2. Comparison of false negative rates across breast density and BI-RADS categories for the baseline model and MM-DETR. MM-DETR shows reduced false negatives in DENSITY B and C, and in BI-RADS 3 and 5, indicating improved sensitivity in diagnostically challenging cases.

false negative rate by 3.7% for DENSITY B and 5.9% for DENSITY C, while the rate for DENSITY D remained unchanged. In the BI-RADS analysis, false negatives dropped by 7.3% in BI-RADS 3 and 15.4% in BI-RADS 5, with a minor increase observed in BI-RADS 4. These results indicate improved robustness in detecting lesions that are often subtle or obscured.

To further illustrate MM-DETR's detection performance, Fig. 3 presents qualitative examples of successful detections across BI-RADS categories 3–5 from breast density C. In (a), the single-view model produces a misaligned and undersized box in the CC view, while MM-DETR accurately localizes the mass in both projections. In (b), one lesion in the CC view is missed by both models, while in the MLO view, the single-view model fails to detect one of the smaller lesions that MM-DETR successfully localizes. In (c), the single-view model produces a noticeably misaligned and undersized prediction in the CC view, whereas MM-DETR achieves a more accurate and well-aligned localization. In the MLO view, both models successfully detect the lesion. These examples highlight the added diagnostic value of multi-view fusion, especially for subtle or complex lesion appearances that may be missed when relying on a single projection.

4.1 Ablation Studies

We continue to explore various fusion methodologies to determine the necessity of bidirectional cross-attention fusion. Table 2 shows results when we fuse the embeddings through dot product, concatenation, cross-attention or bidirectional cross-attention. These results indicate that bidirectional cross-attention is both conceptually sound and empirically the most effective fusion strategy. This is expected, as mutual attention between views enriches contextual information exchange, enhancing the model's ability to detect subtle lesions across projections.

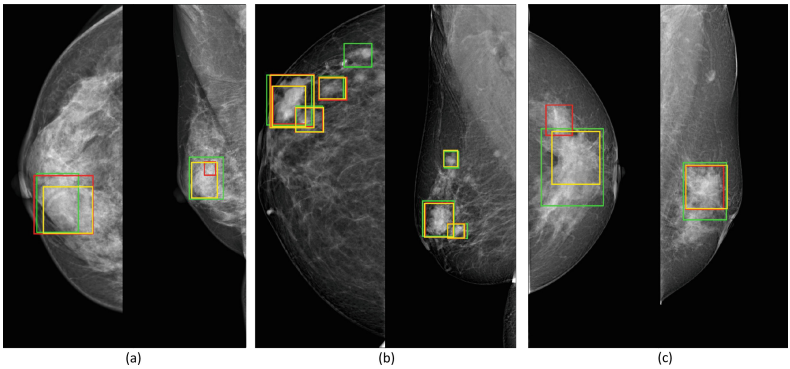


Fig. 3. Qualitative examples of correctly detected masses using MM-DETR across BI-RADS categories 3, 4, and 5. For each case, the CC view is shown on the left and the MLO view on the right. All examples are from DENSITY C cases. Green boxes indicate the ground truth, red boxes show predictions from Grounding DINO (single-view), and yellow boxes represent predictions from MM-DETR. Fig(s) are resized for presentation only. (Color figure online)

Table 2. Comparison of fusion strategies for mass detection.

Fusion Strategy	Mass (mAP)
Dot Product	0.631 ± 0.007
Concatenation	0.644 ± 0.004
Cross Attention	0.643 ± 0.002
Bidirectional Cross Attention	**0.654 ± 0.008**

5 Conclusion

This study shows that combining both mammographic views enhances contextual information, closely reflecting clinical workflow. Towards this end, we introduce MM-DETR, a multi-view, multi-modal detector, achieving a 12.8% and 5.0% improvement over previous SOTA methods. We also evaluate MM-DETR's effectiveness in dense breast tissue, where detection is especially challenging. The model lowers the false negative rate in DENSITY C cases by 5.9%, demonstrating greater reliability in challenging situations. Lastly, we assessed various fusion strategies and found that bidirectional cross-attention delivered the best results by allowing a symmetrical exchange of information between both views.

Moving forward, we plan to extend our multi-view framework beyond mammography to other medical imaging domains, such as Chest X-Rays (CXRs). CXRs often include frontal and lateral views, with some findings more easily diagnosed in one view than the other. We also aim to study alternative fusion locations to enhance cross-view integration. Consequently, we aim to improve

detection by modeling human-like reasoning through multi-view processing and cross-modal integration.

References

1. Bray, F., et al.: Global cancer statistics 2022: globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J. Clin. **74**(3), 229–263 (2024)
2. Giaquinto, A.N., et al.: Breast cancer statistics 2024. CA Cancer J. Clin. **74**(6), 477–495 (2024)
3. Wang, L.: Mammography with deep learning for breast cancer detection. Front. Oncol. (2024)
4. Sickles, E.: Mammographic features of 300 consecutive nonpalpable breast cancers. Am. J. Roentgenol. **146**(4), 661–663 (1986)
5. Kerlikowske, K., et al.: Impact of BMI on prevalence of dense breasts by race and ethnicity. Cancer Epidemiol. Biomark. Prev. **32**(11), 1524–1530 (2023)
6. Boyd, N.F., et al.: Mammographic density and the risk and detection of breast cancer. N. Engl. J. Med. **356**(3), 227–236 (2007)
7. Abdelrahman, L., Al Ghamdi, M., Collado-Mesa, F., Abdel-Mottaleb, M.: Convolutional neural networks for breast cancer detection in mammography: a survey. CBM **131**, 104248 (2021)
8. Agarwal, R., Diaz, O., Lladó, X., Yap, M.H., Martí, R.: Automatic mass detection in mammograms using deep convolutional neural networks. JMI **6**(3), 031409 (2019)
9. Xi, P., Shu, C., Goubran, R.: Abnormality detection in mammography using deep convolutional neural networks. In: 2018 IEEE International Symposium on Medical Measurements and Applications (MeMeA), pp. 1–6 (2018)
10. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks (2016)
11. Jocher, G., Qiu, J., Chaurasia, A.: Ultralytics YOLOv8. Ultralytics (2023)
12. Cao, Z., et al.: DeepLIMa: deep learning based lesion identification in mammograms. In: 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), Seoul, Korea (South), October 2019, pp. 362–370. IEEE (2019)
13. Mahoro, E., Akhloufi, M.A.: Breast masses detection on mammograms using recent one-shot deep object detectors. In: 2023 5th International Conference on Bioengineering for Smart Technologies (BioSMART), pp. 1–4 (2023)
14. Abdelhafiz, D., Yang, C., Ammar, R., Nabavi, S.: Deep convolutional neural networks for mammography: advances, challenges and applications. BMC Bioinform. **20**(S11), 281 (2019)
15. Yang, Z., et al.: MommiNet-v2: mammographic multi-view mass identification networks. Med. Image Anal. **73**, 102204 (2021)
16. Kebede, S.R., et al.: Dual view deep learning for enhanced breast cancer screening using mammography. Sci. Rep. **14**(1), 3839 (2024)
17. Li, L.H., et al.: Grounded Language-Image Pre-training (2022)
18. Liu, S., et al.: Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection (2024)
19. Wang, Z., Zhenbang, W., Agarwal, D., Sun, J.: MedCLIP: Contrastive Learning from Unpaired Medical Images and Text (2022)

20. Chaoyi, W., Zhang, X., Zhang, Y., Wang, Y., Xie, W.: MedKLIP: Medical Knowledge Enhanced Language-Image Pre-Training in Radiology (2023)
21. Ghosh, S., Poynton, C.B., Visweswaran, S., Batmanghelich, K.: Mammo-CLIP: a vision language foundation model to enhance data efficiency and robustness in mammography. In: MICCAI 2024, vol. 15012, pp. 632–642. Springer, Cham (2024)
22. Nguyen, H.T., et al.: VinDr-Mammo: a large-scale benchmark dataset for computer-aided diagnosis in full-field digital mammography (2022)
23. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-End Object Detection with Transformers (2020)
24. Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal Loss for Dense Object Detection (2017)
25. Rezatofighi, H., Tsoi, N., Gwak, J.Y., Sadeghian, A., Reid, I., Savarese, S.: Generalized Intersection over Union: A Metric and A Loss for Bounding Box Regression (2019)
26. Carmen, M.G., et al.: Mammographic breast density and race. Am. J. Roentgenol. **188**(4), 1147–1150 (2007)
27. MMDetection Contributors. OpenMMLab detection toolbox and benchmark (2018)
28. Alsentzer, E., et al.: Publicly Available Clinical BERT Embeddings (2019)
29. Lin, T.-Y., et al.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48
30. Loshchilov, I., Hutter, F.: Decoupled Weight Decay Regularization (2017)
31. Redmon, J., Farhadi, A.: YOLOv3: An Incremental Improvement (2018)



Predicting Aesthetic Outcomes of Breast Cancer Surgery: A Robust and Explainable Image Retrieval Approach

Pedro Ferreira[1(✉)], Mohammad Hossein Zolfagharnasab[1], Tiago Gonçalves[1], Eduard Bonci[2], Carlos Mavioso[2], Maria J. Cardoso[2], and Jaime S. Cardoso[1]

[1] INESC TEC, Faculdade de Engenharia, Universidade do Porto, Rua Dr. Roberto Frias, 4200-465 Porto, Portugal
{pedro.ferreira,mohammad.h.zolfagharnasab}@inesctec.pt
[2] Champalimaud Foundation, Avenida Brasília, 1400-038 Lisboa, Portugal

Abstract. Accurate retrieval of post-surgical images plays a critical role in surgical planning for breast cancer patients. However, current content-based image retrieval methods face challenges related to limited interpretability, poor robustness to image noise, and reduced generalization across clinical settings. To address these limitations, we propose a multistage retrieval pipeline integrating saliency-based explainability, noise-reducing image pre-processing, and ensemble learning. Evaluated on a dataset of post-operative breast cancer patient images, our approach achieves contrastive accuracy of 77.67% for Excellent/Good and 84.98% for Fair/Poor outcomes, surpassing prior studies by 8.37% and 11.80%, respectively. Explainability analysis provided essential insight by showing that feature extractors often attend to irrelevant regions, thereby motivating targeted input refinement. Ablations show that expanded bounding box inputs improve performance over original images, with gains of 0.78% and 0.65% contrastive accuracy for Excellent/Good and Fair/Poor, respectively. In contrast, the use of segmented images leads to a performance drop (1.33% and 1.65%) due to the loss of contextual cues. Furthermore, ensemble learning yielded additional gains of 0.89% and 3.60% over the best-performing single-model baselines. These findings underscore the importance of targeted input refinement and ensemble integration for robust and generalizable image retrieval systems.

Keywords: Medical Image Retrieval Systems · Predicting Surgery Outcome · Breast Cancer · Ensemble Learning · Explainable Retrieval

1 Introduction

Breast cancer is the most common cancer among women, with an estimated 2.3 million new cases each year [1]. To help patients understand possible surgical outcomes, clinicians currently rely on the manual selection of visually similar post-surgical outcomes. However, this approach is time-consuming and error-prone, especially when dealing with large databases [2,3].

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
T. Zhang et al. (Eds.): Deep-Breath 2025, LNCS 16142, pp. 268–278, 2026.
https://doi.org/10.1007/978-3-032-05559-0_27

Content-based image retrieval (CBIR) offers a promising solution to this problem. However, despite its potential, CBIR often faces several challenges, such as low interpretability [4], limited generalization across clinical settings [5] (e.g., differences in zooming), and a lack of robustness to image noise [6] (e.g., presence of background artifacts).

Building on the framework of Zolfagharnasab *et al.* [7], this work aims to analyze the behavior of retrieval models using explainable saliency maps, improve model generalization, and enhance robustness to visual noise and irrelevant features, ultimately improving retrieval performance. Specifically, the main contributions of this work are as follows:

1. A comparative evaluation of state-of-the-art (SoTA) architectures for CBIR in breast cancer aesthetic prediction outcomes;
2. Application of explainable AI (XAI) through visual saliency maps to assess whether retrieval models focus on anatomically relevant image regions;
3. Integration of breast segmented images and bounding box (BB) strategies to ensure retrieval's focus on anatomically relevant image regions;
4. Development and analysis of ensemble learning strategies to combine multiple feature extraction models.

The code related to the implementation of this paper is publicly available in a GitHub repository[1].

2 Related Work

To contextualize our methodology within the broader research landscape, we structure the related work into three areas relevant to our study.

Feature Extractors. Recent research in image retrieval systems focused heavily on improving feature extraction techniques [8]. For this purpose, convolutional neural networks (CNNs) have been widely used. However, recent studies have suggested that Vision Transformers (ViTs) offer superior performance. For instance, Lo *et al.* [9] demonstrated that Google ViT outperformed DenseNet201 in detecting microsatellite instability in colorectal cancer histology. Similarly, Pundhir *et al.* [10] showed that Swin Transformer achieved superior performance to CNN architectures (i.e., DenseNet121, ResNet101, and VGG16) in retrieving similar melanoma images. Recently, a high CBIR performance was obtained by Dordevic and Kumar [11] using GC ViT, a model that effectively captures both local and global features.

Interpretability. Despite CBIR advances, clinical acceptance remains limited largely due to clinicians' lack of trust in AI systems [12]. To address this problem, several XAI techniques have been proposed. Namely, Vasu *et al.* [13] incorporated XAI saliency maps into a CBIR system and demonstrated that visual explanations significantly improved users' ability to interpret results and make

[1] https://github.com/Pedrofs9/cbir-ensemble-xai-segmentation.

informed decisions compared to a non-XAI baseline. Additional work, such as Silva *et al.* [14], demonstrated how visual explanations can improve transparency in X-ray retrieval tasks.

CBIR Improvement. Beyond XAI, various strategies have been proposed to improve CBIR performance and robustness. These include multimodal inputs [7], saliency-guided feature selection [15], and pre-processing techniques to reduce noise [16]. Specifically, in a study conducted by Chadha *et al.* [17], region-of-interest extraction via image cropping improved performance.

3 Methodology

This study proposes a multi-step CBIR pipeline composed of four core modules: (1) region-based pre-processing, (2) feature extraction, (3) explainability, and (4) ensemble, as illustrated in Fig. 1. The dataset and annotation protocol follow the methodology described in the reference work [7]. For each query image, two catalogs were created–one with patients showing Excellent/Good aesthetic surgery outcomes and another with Fair/Poor, based on the Harris scale [18]. Clinicians ranked 515 images per catalog from most to least similar. These rankings were used to construct triplets with an anchor (query), a positive (similar to the query), and a negative (dissimilar to the query) images. The individual feature extractors and ensemble models were trained with this triplet-loss formulation to enforce semantic alignment with clinician-provided similarity rankings. Similarity between the query and retrieved samples was computed using Euclidean distance.

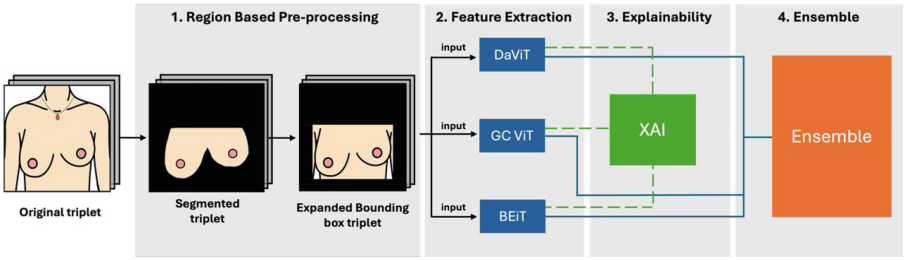


Fig. 1. Scheme of the proposed CBIR pipeline.

To identify the most effective feature extractors for our pipeline, ablation studies were conducted across several SoTA architectures. Furthermore, to assess retrieval models' behavior and improve interpretability, various XAI techniques were tested, ensuring compatibility with CNN and ViT architectures.

To evaluate the impact of visual noise removal, ablation studies were performed using different input strategies. Using the best-performing individual feature extractors, four types of image inputs were compared: original images, segmented breasts, tight BB around the breast region, and expanded bounding box (EBB) versions. The BB was defined as the smallest axis-aligned rectangle

enclosing all pixels labeled as either breast or nipple, while the EBB extended this region by 10 pixels horizontally and 20 vertically to preserve additional contextual information. Finally, to increase robustness by combining the complementary strengths of multiple models, ensemble learning was explored. The effect of multiple fusion types and training strategies was evaluated.

Retrieval performance was assessed using the same metrics as in work [7]: Contrastive Accuracy (CAcc), which measures the ability to identify the more similar image between two options, and Adjusted Discounted Cumulative Gain (ADCG), which evaluates the retrieval relevance and ranking quality.

4 Results and Discussion

To structure the evaluation, the results and discussion are organized into a series of ablation studies comprising each section of the proposed pipeline, followed by the final pipeline assessment. This section is divided into five parts: (1) feature extractors, (2) XAI analysis, (3) limiting exposure to noise, (4) ensemble learning, and (5) final pipeline evaluation.

Feature Extractors. To identify the most effective feature extractor for our pipeline, we evaluated several SoTA CNN and ViT architectures, with fine-tuned learning rates and training until performance convergence. These results are presented in Table 1. Among CNNs, ResNeXt50 performed best, likely due to its grouped convolutions (cardinality), which improves feature representation [19]. Nevertheless, ViTs consistently outperformed CNN architectures, confirming the importance of modeling long-range dependencies.

Table 1. Retrieval performance of CNN and ViT models. Values reported by aesthetic quality (Excellent/Good, Fair/Poor). Best results per metric and architecture are highlighted in bold.

Architecture	Model	CAcc (%) ↑	ADCG (%) ↑
CNN	DenseNet121 [20]	(70.66, 72.68)	(73.14, 78.36)
	DenseNet201 [20]	(69.21, 73.70)	(70.44, 77.19)
	ResNeXt50 [21]	(**72.67**, **73.79**)	(**75.69**, **79.21**)
	VGG16 [22]	(72.34, 73.32)	(75.15, 75.21)
ViT	BEiT [23]	(74.44, 74.78)	(78.33, 78.51)
	DaViT [24]	(74.56, 75.44)	(**78.56**, 79.90)
	GC ViT [25]	(**75.11**, **76.65**)	(78.44, **80.18**)
	Google ViT [26]	(73.56, 73.89)	(78.06, 75.35)
	SwinV2 [27]	(72.67, 75.44)	(75.54, 79.14)

GC ViT achieved the highest CAcc across both Excellent/Good and Fair/Poor sets (75.11% and 76.65%, respectively), followed closely by DaViT

(74.56% and 75.44%). For ADCG, GC ViT and DaViT also achieved the highest results. The hybrid designs of GC ViT and DaViT, which balance local and global representations, outperformed models focusing on capturing solely global (e.g., BEiT) or local dependencies (e.g., VGG16).

XAI Analysis. Despite strong overall results, individual models' performance plateaued below 77% CAcc and 81% ADCG, motivating a qualitative inspection of the retrieved results. Although most of the rankings were visually accurate, such as the one in Fig. 2(a), some had more frequent errors that suggested that the models may not have focused solely on the clinically relevant region, the breasts (Fig. 2(b)).

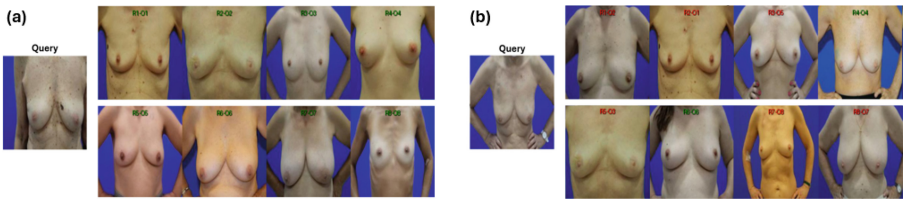


Fig. 2. Image retrieval results for GC ViT. The left image is the query (pre-surgical); the right images are the retrieved post-surgical ones, ordered from left to right by their predicted similarity to the query. (a) Example of a perfect ranking; (b) Example of a ranking with several ordering errors. Image labels follow the format RxOy, where Rx indicates the retrieval rank and Oy denotes the original (ground-truth) similarity rank.

In this ranking, the most evident error occurs between the third and fifth retrieved images. The similarity between the arms and hands placement of the third retrieved image suggests that the retrieval model might have wrongly focused on those regions.

To verify this claim, integrated gradients (IG) [28] saliency maps were generated on that ranking (Fig. 3(a)), confirming that the retrieval models also focused on irrelevant regions such as the arms' position. To verify the consistency of these results with different retrieval models, IG was also applied to BEiT's retrieved rankings, presented in Fig. 3(b). Furthermore, other combinations of retrieval models and XAI methods were also tested: Similarity-Based Saliency Maps (SBSM) [29] for BEiT (Fig. 3(c) and Class Activation Maps (CAM) [30] for DenseNet121 (Fig. 3(d)).

These visual saliency maps corroborate our hypothesis that, although the retrieval models gave greater attention to the breast area, considerable attention was also present in non-breast tissue.

Limiting Exposure to Noise. Based on the XAI results, several input strategies were evaluated to remove visual noise and guide model attention. As shown in Table 2, segmentation generally degraded retrieval performance, likely due to the loss of important contextual cues. In contrast, BB preserved these cues, improving upon the results obtained with segmentation. The best results were achieved with EBB, which outperformed the original images in most cases.

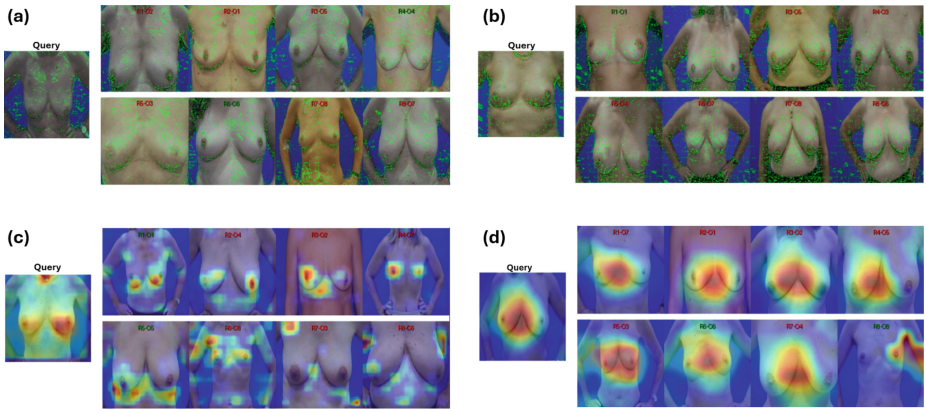


Fig. 3. Image retrieval rankings with XAI attribution maps. The left image is the query (pre-surgical); the right images are retrieved post-surgical ones, ordered by predicted similarity. Saliency maps overlay retrieved images to highlight regions influencing similarity predictions. (a) IG with GC ViT; (b) IG with BEiT; (c) SBSM with BEiT; (d) CAM with DenseNet121. Image labels follow the format RxOy, where Rx indicates the retrieval rank and Oy denotes the original (ground-truth) similarity rank.

An exception was observed with BEiT on the Excellent/Good catalogs, where performance declined relative to the original images. This result likely stems from BEiT's reliance on global attention mechanisms, without localized feature extraction, making it more sensitive to context removal. Notably, this trend was not observed on the Fair/Poor catalogs, where global context is presumably less important due to more frequent asymmetries and scarring.

These results confirm that removing irrelevant image regions, while preserving contextual cues, enhances the model's ability to focus on relevant features, thereby improving retrieval performance.

Table 2. Retrieval performance across various input types. Values reported by aesthetic quality (Excellent/Good, Fair/Poor). Best results per metric are highlighted in bold.

Input Image Type	BEiT		DaViT		GC ViT	
	CAcc (%) ↑	ADCG (%) ↑	CAcc (%) ↑	ADCG (%) ↑	CAcc (%) ↑	ADCG (%) ↑
Original	(**74.44**, 74.34)	(**78.33**, 74.36)	(74.56, 75.44)	(78.56, **79.90**)	(75.11, 76.65)	(78.44, **80.18**)
Segmented	(67.89, 70.72)	(73.04, 73.37)	(72.80, 74.05)	(76.75, 76.30)	(73.78, 75.00)	(**81.35**, 74.29)
BB	(70.78, 72.81)	(74.62, 75.10)	(75.50, 75.60)	(79.60, 79.20)	(75.00, 76.43)	(81.27, 79.95)
EBB	(70.78, **75.88**)	(73.93, **77.22**)	(**75.89**, **75.76**)	(**79.97**, 79.20)	(**75.89**, **77.30**)	(81.17, 79.80)

Ensemble Learning. To further improve retrieval robustness by leveraging diverse architectural strengths, the four best-performing individual models were combined into an ensemble. Ablation studies tested several fusion strategies of increasing complexity: raw feature concatenation, a Single-Layer Projection (SLP) using one fully connected layer followed by ReLU, Multi-Layer Perceptrons (MLPs) followed by ReLU, and a transformer-based decoder. Each strategy

was evaluated in two training regimes: frozen backbone weights and a fully trainable two-phase fine-tuning approach where the fusion layers are trained alone for the first 30% of epochs, followed by joint training of both backbone models and fusion layers for the remaining 70%. The results are presented in Table 3.

Table 3. Retrieval performance comparison between frozen and trainable base models and different feature fusion types. Values reported by aesthetic quality (Excellent/Good, Fair/Poor). Best results per metric are highlighted in bold.

Fusion Type	Frozen Models		Trainable Models	
	CAcc (%) ↑	ADCG (%) ↑	CAcc (%)	ADCG (%)
Raw Concatenation	(71.78, 71.48)	(71.04, 72.02)	N/A	N/A
SLP	(72.44, **72.98**)	(**77.68**, 77.34)	(**76.00**, **80.25**)	(78.93, 78.24)
MLP (2 layers)	(**73.11**, 72.94)	(76.29, **77.45**)	(75.89, 79.31)	(**79.58**, **81.50**)
MLP (3 layers)	(71.82, 72.24)	(74.30, 74.43)	(72.92, 75.14)	(76.54, 77.87)
Transformer	(72.33, 72.55)	(75.32, 75.77)	(74.56, 76.33)	(78.34, 79.11)

Ensembles with trainable base models consistently outperformed frozen ones across all settings, confirming the benefit of end-to-end training. Among fusion types, raw concatenation performed worst, as expected, due to the lack of trainable parameters capable of capturing meaningful feature relationships. In contrast, SLP and two-layer MLP fusions achieved the best performances. The former provides a simple linear combination, while the latter introduces mild nonlinearity to better capture more complex inter-model relationships. These results suggest that mild nonlinearity suffices for effective fusion. On the other hand, deeper MLPs and transformer-based fusion underperformed, likely due to overfitting caused by the added complexity on a limited dataset size. Lastly, using the best ensemble configuration based on CAcc (SLP with trainable base models), the influence of combining different numbers of models was evaluated.

The highest performance was achieved using an ensemble composed of the three top-performing models (BEiT, DaViT, and GC ViT), obtaining 77.67% CAcc and 80.31% ADCG on Excellent/Good, and 80.14% CAcc and 80.63% ADCG on the Fair/Poor. This ensemble achieved the most effective balance between model diversity and performance-driven model selection.

Final Pipeline. The complete pipeline combined the best ensemble setups from the ablation studies (SLP or two-layer MLP with top-3 models) with EBB inputs. Additionally, we also evaluated ensembles using the top-2 models, excluding BEiT due to its subpar performance on the Excellent/Good catalogs set. These results are presented in Table 4.

Unlike earlier results (see Table 3), in which the SLP and two-layer MLP yielded comparable performance (with a slight advantage for the SLP in CAcc), the EBB setting revealed a clear superiority of the two-layer MLP. It achieved 77.67% CAcc on Excellent/Good and 84.98% on Fair/Poor, outperforming the

Table 4. Retrieval performance of ensemble variations using EBB images. Values reported by aesthetic quality (Excellent/Good, Fair/Poor). Best results per metric are highlighted in bold.

Fusion Type	Models	CAcc (%) ↑	ADCG (%) ↑
SLP	DaViT+GC ViT	(77.00, 83.11)	(80.38, 74.63)
	BEiT+DaViT+GC ViT	(75.33, 81.80)	(80.00, 76.30)
MLP (2 layers)	DaViT+GC ViT	(77.44, 83.11)	(**82.98**, 74.63)
	BEiT+DaViT+GC ViT	(**77.67**, **84.98**)	(82.55, **78.64**)

top-3 models SLP, which reached 75.33% and 81.80%, respectively. The best-performing ensemble, showed a clear advantage over the best individual model, GC ViT (see Table 2), which had obtained 75.89% CAcc on Excellent/Good and 77.30% on Fair/Poor, underscoring the value of ensemble learning.

Compared to the best ensemble using original images (see Table 3), the EBB configuration yielded superior performance: while both achieved 77.67% CAcc on Excellent/Good, the expanded version reached 84.98% on Fair/Poor, compared to 80.14% with original images. Similarly, ADCG on Excellent/Good improved from 80.31% to 82.55%, although Fair/Poor dropped from 80.63% to 78.64%, likely due to higher variability in these catalogs (more scarring and asymmetries), where combining multiple models may dilute critical visual cues. These results confirm that removing irrelevant image regions, while preserving contextual cues, enhances the model's ability to focus on relevant features, consequently improving retrieval performance. Furthermore, combining diverse architectures through ensemble learning has proven to further improve overall performance.

Table 5 compares our proposed pipeline with the image-only approach from the previous study [7]. Our method achieved an 8.37% increase in CAcc and 8.68% in ADCG for Excellent/Good outcomes. For Fair/Poor outcomes, CAcc improved by 11.80%, while ADCG showed a decline of 1.22%. These substantial gains across most metrics highlight the effectiveness of our approach.

Table 5. Comparison of retrieval performance between the reference approach [7] and the proposed pipeline. Values reported by aesthetic quality (Excellent/Good, Fair/Poor). Best results per metric are highlighted in bold.

Method	CAcc (%) ↑	ADCG (%) ↑
Zolfagharnasab *et al.* [7]	(69.30, 73.18)	(73.87, **79.86**)
Proposed Pipeline	(**77.67**, **84.98**)	(**82.55**, 78.64)

5 Conclusions and Future Work

The proposed pipeline, combining region-based image pre-processing with ensemble learning, led to notable improvements in noise robustness and model

generalization. Compared to the previous study [7], it achieved an 8.37% increase in CAcc for Excellent/Good outcomes and 11.80% for Fair/Poor. By leveraging XAI saliency maps to explore model performance, the use of EBB inputs led to gains of 0.78% and 0.65% CAcc for Excellent/Good and Fair/Poor outcomes, respectively, proving the importance of context-preserving noise reduction. Additionally, combining diverse architectures through ensemble learning led to 0.89% and 3.60% increases over the best-performing single-model baselines.

Future work could explore dynamically selecting the most informative image regions. Furthermore, alternative methods such as collaborative filtering can be explored for reducing reliance on labels or as retrieval-augmented generation (RAG) for providing more automated capturing of similarity measures as well as multimodal and visual and textual support.

Acknowledgements. This work has received funding from the European Union's Horizon Europe research and innovation programme under the Grant Agreement 101057389-CINDERELLA project, and by the Portuguese Foundation for Science and Technology (FCT) through the Ph.D. Grant "2024.06378.BD".

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

Data Policy. The dataset used for this study is private, but is available upon request.

References

1. Lukasiewicz, S., Czeczelewski, M., Forma, A., Baj, J., Sitarz, R., Stanislawek, A.: Breast cancer-epidemiology, risk factors, classification, prognostic markers, and current treatment strategies-an updated review. Cancers **13**(17), 4287 (2021)
2. Silva, W., Carvalho, M., Mavioso, C., Cardoso, M.J., Cardoso, J.S.: Deep aesthetic assessment and retrieval of breast cancer treatment outcomes. In: Pattern Recognition and Image Analysis, 10th Iberian Conference, pp. 108–118 (2022)
3. Zolfagharnasab, M.H., PourMohammadBagher, L., Bahrani, M.: Intelligent travel recommendations using neural collaborative filtering for touristic landmarks of Iran. J. Data Sci. Model., 119–147 (2025)
4. Alzu'bi, A., Amira, A., Ramzan, N.: Semantic content-based image retrieval: a comprehensive study. J. Vis. Commun. Image Represent. **32**, 20–54 (2015)
5. Wan, J., et al.: Deep learning for content-based image retrieval: a comprehensive study. In: Proceedings of the 22nd ACM International Conference on Multimedia, pp. 157–166 (2014)
6. Bilquees, S., Dawood, H., Dawood, H., Majeed, N., Javed, A., Mahmood, M.T.: Noise resilient local gradient orientation for content-based image retrieval. Int. J. Opt. **2021**, 4151482 (2021)
7. Zolfagharnasab, M.H., et al.: Predicting aesthetic outcomes in breast cancer surgery: a multimodal retrieval approach. In: Deep Breast Workshop on AI and Imaging for Diagnostic and Treatment Challenges in Breast Care, pp. 137–147. Springer (2024)

8. Shiv Ram Dubey: A decade survey of content based image retrieval using deep learning. IEEE Trans. Circuits Syst. Video Technol. **32**(5), 2687–2704 (2022)
9. Lo, C.-M., Jiang, J.-K., Lin, C.-C.: Detecting microsatellite instability in colorectal cancer using transformer-based colonoscopy image classification and retrieval. PLoS ONE **19**(1), 1–16 (2024)
10. Pundhir, A., Sagar, S., Singh, P., Raman, B.: Echoes of images: multi-loss network for image retrieval in vision transformers. Med. Biol. Eng. Comput. **62**, 1–15 (2024)
11. Dordevic, D., Kumar, S.: Evidential transformers for improved image retrieval. *arXiv preprint*, arXiv:2409.01082 (2024)
12. Sadeghi, Z., et al.: A review of explainable artificial intelligence in healthcare. Comput. Electr. Eng. **118**, 109370 (2024)
13. Vasu, B., Brian, H., Dong, B., Collins, R., Hoogs, A.: Explainable, interactive content-based image retrieval. Appl. AI Lett. **2**(4), e41 (2021)
14. Silva, W., et al.: Computer-aided diagnosis through medical image retrieval in radiology. Sci. Rep. **12**(1), 20732 (2022)
15. Gassner, M., et al.: Saliency-enhanced content-based image retrieval for diagnosis support in dermatology consultation. JMIR Dermatol. **6**, e42129 (2023)
16. Ahmad, F.: Deep image retrieval using artificial neural network interpolation and indexing based on similarity measurement. CAAI Trans. Intell. Technol. **7**(2), 200–218 (2022)
17. Chadha, A., Mallik, S., Johar, R.: Comparative study and optimization of feature-extraction techniques for content based image retrieval. Int. J. Comput. Appl. **52**(17), 1–9 (2012)
18. Rose, M.A., et al.: Conservative surgery and radiation therapy for early breast cancer: long-term cosmetic results. Arch. Surg. **124**(2), 153–157 (1989)
19. Zolfagharnasab, M.H., Bahrani, M., Hamed Saghayan, M., Masoumi, F.S.: Exploring a novel multi-channel structure to improve facial expression recognition on occluded samples using deep convolutional neural network. J. Artif. Intell. Appl. Innov. **1**(2), 26–41 (2024)
20. Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. *arXiv preprint*, arXiv:1608.06993 (2018)
21. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5987–5995 (2017)
22. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. *arXiv preprint*, arXiv:1409.1556 (2015)
23. Bao, H., Dong, L., Piao, S., Wei, F.: BEiT: BERT pre-training of image transformers. *arXiv preprint*, arXiv:2106.08254 (2022)
24. Ding, M., Xiao, B., Codella, N., Luo, P., Wang, J., Yuan, L.: DaViT: dual attention vision transformers. *arXiv preprint*, arXiv:2204.03645 (2022)
25. Hatamizadeh, A., Yin, H., Heinrich, G., Kautz, J., Molchanov, P.: Global context vision transformers. *arXiv preprint*, arXiv:2206.09959 (2023)
26. Dosovitskiy, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale. *arXiv preprint*, arXiv:2010.11929 (2021)
27. Liu, Z., et al.: Swin transformer v2: scaling up capacity and resolution. *arXiv preprint*, arXiv:2111.09883 (2022)
28. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: Proceedings of the 34th International Conference on Machine Learning, vol. 70, pp. 3319–3328 (2017)

29. Dong, B., Collins, R., Hoogs, A.: Explainability for content-based image retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 1–13 (2019)
30. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2921–2929 (2016)

Towards Utilizing Robust Radiance Fields for 3D Reconstruction of Breast Aesthetics

Gonçalo Pinto[1(✉)], Mohammad Hossein Zolfagharnasab[1], Luís F. Teixeira[1], Helena Cruz[2,3], Maria J. Cardoso[2], and Jaime S. Cardoso[1]

[1] INESC TEC, Faculdade de Engenharia, Universidade do Porto, Rua Dr. Roberto Frias, 4200-465 Porto, Portugal
up202004907@edu.fe.up.pt, mohammad.h.zolfagharnasab@inesctec.pt
[2] Champalimaud Foundation, Avenida Brasília, 1400-038 Lisboa, Portugal
[3] ISCTE – Instituto Universitário de Lisboa, Avenida das Forças Armadas, 1649-026 Lisboa, Portugal

Abstract. 3D models are crucial in predicting aesthetic outcomes in breast reconstruction, supporting personalized surgical planning, and improving patient communication. In response to this necessity, this is the first application of Radiance Fields to 3D breast reconstruction. Building on this, the work compares six SoTA 3D reconstruction models. It introduces a novel variant tailored to medical contexts: Depth-Splatfacto, designed to improve denoising and geometric consistency through pseudo-depth supervision. Additionally, we extended model training to grayscale, which enhances robustness under grayscale-only input constraints. Experiments on a breast cancer patient dataset demonstrate that Splatfacto consistently outperforms others, delivering the highest reconstruction quality (PSNR 27.11, SSIM 0.942) and the fastest training times (×1.3 faster at 200k iterations). At the same time, the depth-enhanced variant offers an efficient and stable alternative with minimal fidelity loss. The grayscale train improves speed by ×1.6 with a PSNR drop of 0.70. Depth-Splatfacto further improves robustness, reducing PSNR variance by 10% and making images less blurry across test cases. These results establish a foundation for future clinical applications, supporting personalized surgical planning and improved patient-doctor communication.

Keywords: 3D Breast Reconstruction · Neural Radiance Fields · Gaussian Splatting · Depth-Infused Scene Representation · Robust View Synthesis

1 Introduction

Reconstructing accurate and photorealistic 3D models of the human body is becoming increasingly vital in medical imaging, especially within plastic and reconstructive surgery [1]. In breast reconstruction, 3D modeling plays a critical role not only in visualizing anatomy but also in predicting aesthetic outcomes–an essential factor in patient satisfaction and psychological recovery [2].

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
T. Zhang et al. (Eds.): Deep-Breath 2025, LNCS 16142, pp. 279–288, 2026.
https://doi.org/10.1007/978-3-032-05559-0_28

These models offer valuable support for surgical planning, patient-doctor communication, and managing patient expectations more effectively [3].

Despite their clinical value, high-fidelity 3D reconstruction tools are often costly and require specialized equipment, limiting their use in many healthcare settings. Meanwhile, traditional 2D photographs, though inexpensive, lack spatial depth, making it difficult to evaluate breast volume, symmetry, and contour, reducing clinical reliability and limiting their usefulness in patient consultations. To address both the technical and practical challenges, this work explores a cost-effective, scalable approach to 3D reconstruction for breast aesthetic evaluation. The proposed method enhances visualization while remaining mindful of real-world clinical constraints. Specifically, this study is:

1. Presenting the first comparative study of state-of-the-art (SoTA) 3D reconstruction methods based on diverse model paradigms, evaluating their performance, robustness, and suitability for clinical breast imaging.
2. Proposing a novel depth-supervised variant of Splatfacto (**Depth-Splatfacto**), which enhances reconstruction robustness, reduces dependence on required sample counts, and incorporates more geometric cues through depth supervision.
3. Integrating noise-robust techniques–including automated background (BG) removal, depth clamping, and artificial inpainting–to streamline the reconstruction pipeline, reduce manual effort, and improve overall model resilience.

To facilitate reproducibility, the implementation code for this study is publicly available on GitHub.[1]

2 Related Work

This section reviews the SoTA work relevant to our study, organized into four thematic areas as following.

SoTA On 3D Reconstruction. Recent advances in 3D reconstruction have focused on improving accuracy without increasing time consumption. For instance, surface-based methods traditionally require manual masks to delineate geometry accurately; NeuS-facto removes this dependency by proposing a new volume rendering method to train Signed Distance Functions (SDFs) representations, though it struggles with textureless objects [4]; in response, Neuralangelo employs a coarse-to-fine gradient scheme to recover fine detail, although with higher memory usage and longer runtimes [5]. Next, Instant-NGP was proposed to accelerate convergence by leveraging multiresolution hash grids to produce results in minutes, but it fails to capture fine details from hash collisions [6]. Nerfacto builds on this hashing scheme, integrating other proposal networks, to strike a practical balance between speed and rendering quality; however, its scores remain below MipNeRF-360's [7]. In contrast, Splatfacto operates purely on 3D Gaussian Splatting (GS) of high-quality point clouds for real-time performance, requiring the original sparse point clouds to use as seed points [8]. To

[1] https://github.com/gpe0/3D-Reconstruction.

date, no prior work has systematically evaluated SoTA techniques on breast aesthetics data. This makes selecting an appropriate baseline challenging, requiring empirical evaluation to verify its performance, as done in our experiments [9].

3D Medical Applications. While Radiance Fields have made significant progress in 3D scene reconstruction, their application in exterior breast (aesthetics) surfaces remains unexplored. Conventional clinical pipelines still rely on voxel- or mesh-based imaging, which are predominantly obtained with depth cameras or via SLAM-based reconstructions [10]. Recent models like UMedNeRF demonstrate the potential of NeRF-based methods to reconstruct anatomy from sparse 2D views. However, these have focused primarily on interior structures [11]; there are no works on breast aesthetics. In this work, 3D reconstruction is extended to exterior breast surfaces.

Depth Supervision. Recent works on 3D modeling have used depth supervision from RGB-D sensors and learned monocular depth networks. For example, Eigen et al. employed multi-scale Convolutional Neural Networks (CNNs) to predict metric depth from a single image and integrate these estimates into dense reconstruction pipelines [12]. More recently, depth cues have been used to guide volumetric rendering models: Depth-supervised NeRF injects sparse structure-from-motion (SfM) point clouds as a free supervisory signal on each ray's termination distribution, yielding faster convergence and plausible geometry from fewer views [13]. This work extends these ideas by adding depth supervision on GS, integrating a depth-constrained loss to boost surface accuracy and training efficiency.

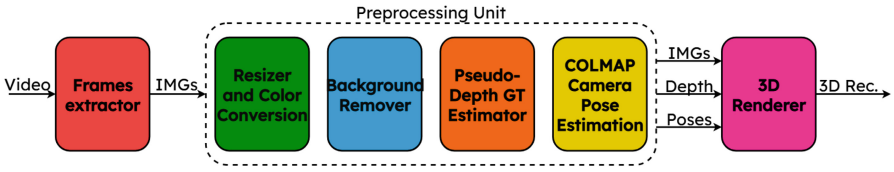


Fig. 1. 3D Model Reconstruction Pipeline.

3 Methodology

In this study, we extended Splatfacto by adding depth supervision to the original radiance network via a multi-head design (depicted in Fig. 1), inspired by Depth-Nerfacto [13].

Rather than supervising along rays, *Depth-Splatfacto (DS)* enforces depth consistency at the image level by comparing the rendered accumulation maps against GT depth using a custom penalization loss as defined in Eq. 1:

$$\mathcal{L}_{\text{total}} = \frac{1}{N}\sum_{i=1}^{N}\left\|\hat{I}_i - I_i\right\|^2 + \lambda_{\text{berHu}} \cdot \begin{cases} |d-\hat{d}|, & \text{if } |d-\hat{d}| \le c \\ \frac{(d-\hat{d})^2+c^2}{2c}, & \text{otherwise} \end{cases} \tag{1}$$

where the first term is the RGB loss, which measures the per-pixel mean squared error between the predicted and GT images, and the second term is the berHu loss, where d is the GT depth, $\hat{d}$ is the predicted depth, and $c = 0.2 \cdot \max(|d - \hat{d}|)$ is a dynamic threshold based on the maximum residual [14].

Regarding the used data loader, the existing *Full-Image-Datamanager* is configured to batch full-resolution RGB images, aligned depth maps corresponding to each input image, and camera poses corresponding to each input image. To stabilize learning, an alpha-clamping heuristic was applied after 500 steps, a value determined empirically through ablation, to suppress artifacts and improve foreground-background separation during training.

With the goal of improved training, we also extended the standard RGB training to one that focuses on geometric changes rather than color differences. To enable grayscale, the spherical-harmonic (SH) radiance representation was collapsed to one channel, reducing the number of parameters without altering pipeline logic. Besides changing the Gaussian representation, the seed-point initialization density was reduced to one-eighth of the original, reducing the reliance on the initial point cloud generated by the SfM, cutting memory usage, and speeding up training. As a result, this adaptation is fully modular and can be applied to any Radiance Field method. It is worth noting that the CUDA-accelerated GSplat kernels were adapted to handle both 1 and 3 channels, thus ending up with a custom GSplat version.

4 Dataset

The dataset used in this study comprises data of the entire torso from a single human breast cancer patient, acquired at the Champalimaud Foundation's experimental platform, which ensures light consistency across the multiple frames. As input, a video sequence was recorded to cover multiple viewpoints of the target, from which 181 frames with a resolution of 1550×741 pixels were extracted. Rather than just using the images, camera poses were estimated using *COLMAP*, to generate sparse point clouds and align camera parameters, important for the training process [15]. Then, we took two additional preprocessing steps to enrich the dataset quality: firstly, to make the models center their rendering power on the right target, *RMBG v1.4*, identified as the strongest candidate in realistically removing BG, was used to create foreground masks [16]; Secondly, in order to penalize depth outputs, *Depth Anything V2* was selected to generate pseudo-depth GT maps from the original images, as it proved to be the top estimator among the tested [17].

To enable accurate 2D visual feedback for aesthetic applications, a lightweight *LAB color fusion* step was applied post-training to reconstruct RGB images. Since PSNR alone does not capture image blurriness, Variance of Laplacian (VoL) was included to assess image sharpness [18].

Finally, all models were implemented using the *Nerfstudio* framework and evaluated with standard metrics (PSNR, SSIM, LPIPS) [19], as well as VoL, on the same machine (NVIDIA RTX A2000, 12 GB GPU) [7].

5 Results and Discussion

We began by systematically comparing SoTA 3D reconstruction methods in the previous detailed human breast cancer patient dataset, quickly revealing Splatfacto as the top performer. Building on that success, we added explicit depth supervision to Splatfacto's original design. As the geometrical cues are more valuable than colorization, we then created and trained a grayscale-only variant, where we only integrated the colorization as the post-processing step of 2D images to maintain realistic colorings when presenting to the patients.

Comparative Analysis. This experiment evaluated six reconstruction models, comparing their performance across perceptual and fidelity metrics and training efficiency. Figure 3 presents the metrics evolution across epochs, while Fig. 2 illustrates qualitative differences in output images. *Splatfacto* was the best-performing model, achieving a PSNR of 27.11, SSIM of 0.942, and the lowest LPIPS (0.163) at 200k epochs, also showing strong early results and efficient training. *Instant-NGP* peaks, then overtrains, *Nerfacto* converges fast at moderate quality, *NeuS* variants improve slowly but lag behind, and *BakedAngelo* performs worst.

On top of this, the charts revealed that volumetric reconstruction methods consistently outperformed surface-based techniques across the different metrics. However, among the surface-based approaches, *NeuS-facto-Angelo* achieved the highest PSNR, distinguishing it as the leading performer within that category. Given the results of this analysis, *Splatfacto* was selected as the foundation for further experiments to explore strategies for robust 3D reconstruction.

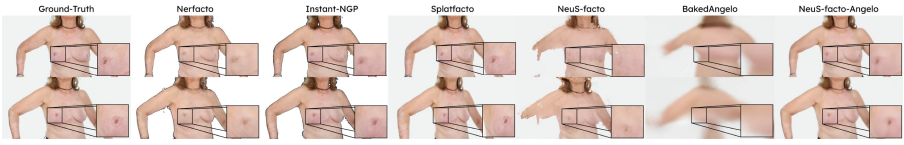


Fig. 2. Visual Comparison of the Comparative Analysis - Showcases the quality of Splatfacto reconstructions compared to the other methods.

Depth Supervision. As 3D modeling is essentially a correct assignment of depth and colorization per location, we investigated the impact of including synthesized depth values as a supplementary signal to correctly align the geometrical cues. To test this, we performed two ablation studies: selecting a depth estimator to get the pseudo-depth GT, and selecting a BG removal model to eliminate irrelevant content.

In this regard, two monocular depth estimators, *Zoe* and *Depth Anything v2*, were used in the proposed method and compared in terms of the standard metrics as shown in Table 1. *Depth Anything v2* outperformed the standard model both in reconstruction quality and in time and therefore was selected.

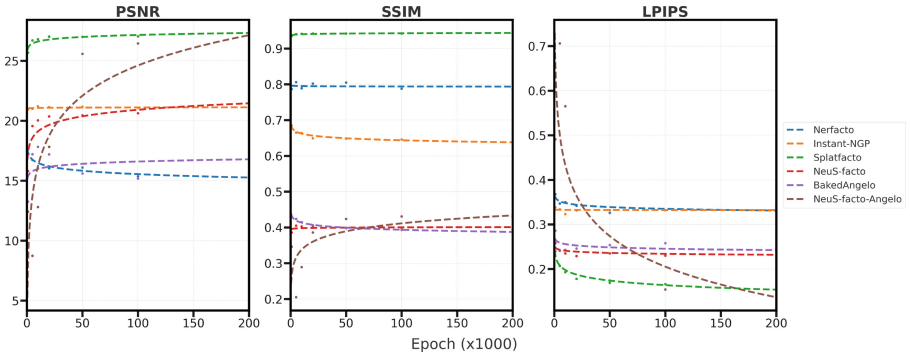


Fig. 3. Comparative Analysis on PSNR ↑, SSIM ↑, and LPIPS ↓ - Splatfacto outperforms every other model and converges faster.

In terms of BG, several segmentation models were evaluated, including: *RMBG v1.4*, *BEN2*, *BiRefNet*, and *Finegrain*. These models were tested in a human sample dataset and are depicted in Table 2. *RMBG v1.4* achieved the highest PSNR and was the second fastest model; however, its segmentation was not perfect, resulting in some artifacts in a few GT images. This error favors Splatfacto as its evaluation considers the entire image, without making BG exclusions, producing blurrier images that blend these anomalies better.

Table 1. Depth Estimators Comparison

Reference	Model	PSNR ↑	SSIM ↑	LPIPS ↓	Training Time (s)
[20]	ZoeDepth	9.59	0.842	0.435	544.72
[17]	Depth Anything v2	**26.41**	**0.939**	**0.192**	**310.55**

Table 2. BG Segmenters Comparison

Reference	Model	Last Update	PSNR ↑	Inference Time (s)
[16]	RMBG v1.4	12 March 2025	**31.82**	0.585
[21]	BEN2	07 April 2025	27.77	**0.448**
[22]	Finegrain	11 September 2024	12.97	13.211
[23]	BiRefNet	31 March 2025	30.31	0.788

As shown in Tables 3 and visually shown in Fig. 4, PSNR dropped slightly, this is due to the fact that *Depth-Splatfacto* divides the image into background and foreground to optimize its training and remove possible artifacts captured in some viewpoints, as illustrated in Fig. 5, whereas our method concurrently stabilizes the learning process and performs denoising during training. Therefore,

some BG pixels still have an effect on it, reducing the overall quality but stabilizing its variation, making the view synthesis more consistent. However, **DS** has shown a substantial improvement when it comes to sharpness, improving from 0.24 to 0.92, meaning the images are way less blurry.

Table 3. Depth Supervision on Splatfacto (100 k iterations); parentheses denote the variation across the different evaluation images

Model	PSNR ↑	SSIM ↑	LPIPS ↓	VoL ↑
Splatfacto	**27.72** (3.07)	**0.944** (0.010)	0.1653 (**0.025**)	0.24
Depth-Splatfacto	26.72 (**2.81**)	0.942 (**0.009**)	**0.1651** (0.026)	**0.92**



Fig. 4. Visual Depth-Splatfacto Results - Portions of the image zoomed on the right, highlighted by red circles. (Color figure online)

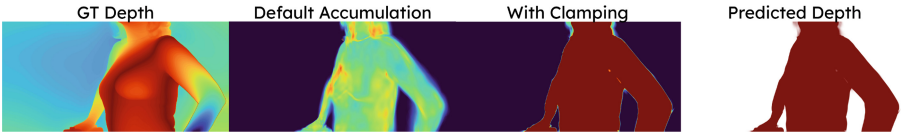


Fig. 5. DS Denoising Capabilities - The default accumulation could not create a good depth prediction; with clamping, the accumulation image is denoised during training, resulting in more accurate depth maps.

As for the grayscale variant, it sped up training by around ×1.6, from 7967 to 4979 s with a small reduction in quality (depicted in Fig. 6). In comparisson, Depth-Splatfacto requires ×1.5 more memory than the original (41.1 MB vs. 27.6 MB at 10k), though the increase is modest given Splatfacto's efficiency.

Sample Reduction. To evaluate if depth supervision compensates for smaller sample sizes, we tested **DS** with Splatfacto on three different scenarios, each containing a smaller number of frames. From the Table 4 results, **Depth-Splatfacto** outperformed its competitor, being more evident as the number of samples reduces, where the images were way less blurry, from 0.08 to 0.72 at 46 frames.

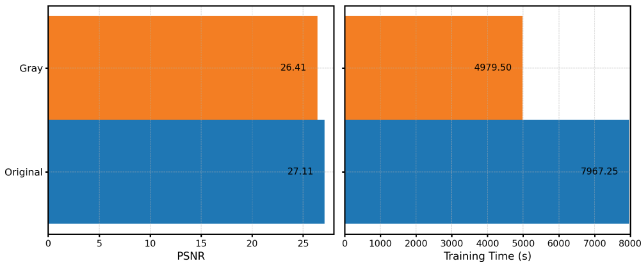


Fig. 6. Effect of training in grayscale compared to the original RGB - Small variation in reconstruction quality with a noticeable speed up in training time.

As for the reconstruction quality, there was a slight decay overall but an improvement in its variation, being a more consistent model across views. You can better confirm these findings in Fig. 7, for example, in the second image at 46 frames, our model completely removes an artifact found in Splatfacto in the breast area.

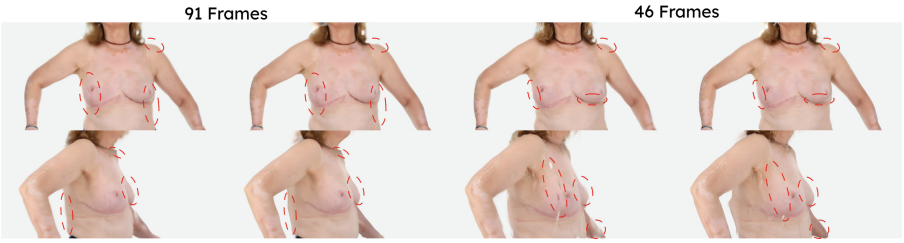


Fig. 7. Sample Reduction Results on 10k iterations (Original, Our Model) - Red circles shows the most evident differences. (Color figure online)

Table 4. Sample Reduction Experiment – 10 k Iterations

Model	91 Frames			46 Frames			23 Frames		
	PSNR ↑	LPIPS ↓	VoL ↑	PSNR ↑	LPIPS ↓	VoL ↑	PSNR ↑	LPIPS ↓	VoL ↑
Splatfacto	**26.95**	0.217	0.03	**26.74**	0.204	0.08	**14.18**	**0.416**	0.10
DepthSplatfacto	25.75	**0.212**	**0.65**	25.48	**0.199**	**0.72**	**14.18**	0.417	**0.90**

6 Conclusions and Future Work

Volumetric rendering methods outperformed surface-based techniques on human-like data, with Splatfacto achieving the best perceptual quality and

fastest convergence, underscoring the suitability of volume-based models for exterior breast surfaces. Our method, **Depth-Splatfacto** has proven to be a more robust model under sparse conditions, creating views that are less blurry and more consistent when compared to the original model. Additionally, the grayscale training accelerates training, contributing to a more robust model by reducing the impact of color noise. To further reduce data dependence in radiance field-based 3D reconstruction, we suggest exploring synthetic viewpoint generation using diffusion models, as it provides faster inference due to pre-trained knowledge as well as reducing the amount of data required for training and promotes generalizability.

Acknowledgements. This work has received funding from the European Union's Horizon Europe research and innovation programme under the Grant Agreement 101057389-CINDERELLA project, and by the Portuguese Foundation for Science and Technology (FCT) through the Ph.D. Grant "2024.06378.BD".

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article

Data Policy. The dataset used for this study is private, but is available upon request.

References

1. Arnold, M., et al.: Current and future burden of breast cancer: global statistics for 2020 and 2040. Breast **66**, 15–23 (2022)
2. Cardoso, J.S., Silva, W., Cardoso, M.J.: Evolution, current challenges, and future possibilities in the objective assessment of aesthetic outcome of breast cancer locoregional treatment. Breast **49**, 123–130 (2020)
3. Zolfagharnasab, M.H., et al.: Predicting aesthetic outcomes in breast cancer surgery: a multimodal retrieval approach. In: Deep Breast Workshop on AI and Imaging for Diagnostic and Treatment Challenges in Breast Care, pp. 137–147. Springer (2024)
4. Wang, P., Liu, L., Liu, Y., Theobalt, C., Komura, T., Wang, W.: Neus: learning neural implicit surfaces by volume rendering for multi-view reconstruction. In: Proceedings of the 35th International Conference on Neural Information Processing Systems, NIPS 2021. Curran Associates Inc., Red Hook (2024)
5. Li, Z., et al.: Neuralangelo: high-fidelity neural surface reconstruction. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8456–8465 (2023)
6. Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding. ACM Trans. Graph. **41**(4), 1–15 (2022)
7. Tancik, M., et al.: Nerfstudio: a modular framework for neural radiance field development. In: ACM SIGGRAPH 2023 Conference Proceedings, SIGGRAPH 2023. Association for Computing Machinery, New York (2023)
8. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3D gaussian splatting for real-time radiance field rendering. ACM Trans. Graph. **42**(4) (2023)

9. Zolfagharnasab, M.H., Bahrani, M., Saghayan, M.H., Masoumi, F.S.: Exploring a novel multi-channel structure to improve facial expression recognition on occluded samples using deep convolutional neural network. J. Artif. Intell. Appl. Innov. **1**(2), 26–41 (2024)
10. Molaei, A., et al.: Implicit neural representation in medical imaging: a comparative survey (2023)
11. Hu, J., Fan, Q., Hu, S., Lyu, S., Wu, X., Wang, X.: Umednerf: uncertainty-aware single view volumetric rendering for medical neural radiance fields. In: 2024 IEEE International Symposium on Biomedical Imaging (ISBI), pp. 1–4 (2024)
12. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. In: Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2, NIPS 2014, pp. 2366–2374. MIT Press, Cambridge (2014)
13. Deng, K., Liu, A., Zhu, J.-Y., Ramanan, D.: Depth-supervised nerf: fewer views and faster training for free. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 12872–12881 (2022)
14. Zhang, H.: Inverse-huber loss based PPO algorithm. In: 2023 5th International Conference on Robotics, Intelligent Control and Artificial Intelligence (RICAI), pp. 986–990 (2023)
15. Schönberger, J.L., Frahm, J.-M.: Structure-from-motion revisited. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4104–4113 (2016)
16. Qin, X., Dai, H., Hu, X., Fan, D.P., Shao, L., Van Gool, L.: Highly accurate dichotomous image segmentation. In: Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, 23–27 October 2022, Proceedings, Part XVIII, pp. 38–56. Springer, Heidelberg (2022)
17. Yang, L., et al.: Depth anything V2 (2024)
18. Bansal, R., Raj, G., Choudhury, T.: Blur image detection using Laplacian operator and Open-CV. In: 2016 International Conference System Modeling & Advancement in Research Trends (SMART), pp. 63–67 (2016)
19. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: representing scenes as neural radiance fields for view synthesis. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) Computer Vision – ECCV 2020, pp. 405–421. Springer, Cham (2020)
20. Bhat, S.F., Birkl, R., Wofk, D., Wonka, P., Muller, M.: Zoedepth: zero-shot transfer by combining relative and metric depth (2023)
21. Meyer, M., Spruyt, J.: Ben: using confidence-guided matting for dichotomous image segmentation (2025)
22. Yu, Q., Zhao, X., Pang, Y., Zhang, L., Lu, H.: Multi-view aggregation network for dichotomous image segmentation (2024)
23. Zheng, P., et al.: Bilateral reference for high-resolution dichotomous image segmentation (2024)

Bridging the Gap: A Community Driven and AI-Enabled Approach to Early Breast Cancer Detection in Black African Women

Adaobi Emegoakor[1], Confidence Raymond[2,3], Yewande Gbadamosi[4], Richard Malumba[5], Charity Umoren[2], Diana Spence Betancourt[6], Aondona M. Iorumbur[2,7], Chinasa Kalaiwo[2], Abbas Rabiu Muhammad[8], Dennis Musinguzi[5], Patience Atukunda[5], Peter Makhoul[9], Rosta Asiimwe[5], Alfred Jatho[10], Amaka Nnamani[11], Franca Eze[12], Oluyemisi Toyobo[2,13], Abiodun Fatade[2,12], Udunna C. Anazodo[2,8], Farouk Dako[2,9], Michael Kawooya[5], and Maruf Adewole[2,9(✉)]

[1] Nnamdi Azikiwe University Teaching Hospital Nnewi, Nnewi, Nigeria
[2] Medical Artificial Intelligence Laboratory (MAI Lab), Lagos, Nigeria
madewole@mailab.io
[3] Department of Biomedical Engineering, McGill University, Montreal, Canada
[4] Lagos State Teaching Hospital, Lagos, Nigeria
[5] Ernest Cook Ultrasound Research and Education Institute (ECUREI), Kampala, Uganda
[6] University of Pennsylvania, Philadelphia, PA, USA
[7] University of Lagos, Lagos, Nigeria
[8] Department of Radiology, Bayero University, Kano, Nigeria
[9] Department of Radiology, University of Pennsylvania, Philadelphia, USA
[10] Uganda Cancer Institute, Kampala, Uganda
[11] University of Nigeria Teaching Hospital, Enugu, Nigeria
[12] My Body My Asset Cancer Foundation, Lagos, Nigeria
[13] Crestview Radiology Ltd., Lagos, Nigeria

Abstract. Breast cancer incidence is rapidly rising in sub-Saharan Africa (SSA), with projections indicating a doubling by 2040. Black African women experience the highest mortality rates globally, primarily due to late-stage diagnosis and limited access to diagnostic services. There is an urgent need for innovative, low-cost, and accessible solutions to improve early detection. We integrated community-based participatory research in a prospective, multi-site imaging study conducted in Nigeria and Uganda. The study aims to develop a machine learning (ML) approach for early detection of breast lesions on point-of-care ultrasound (POCUS). Over 1000 women have participated in community outreaches and received at-home breast screening training and breast cancer prevention (nutrition, exercise, etc.) resources. Currently, a total of 407 women from the community events are enrolled in the study and have received POCUS screening onsite. Breast lesions were identified in 31 women (8.6%), of whom 14 (45%) completed follow-up imaging and clinical evaluations. The POCUS images are being preprocessed and prepared for an imaging challenge to enable development of robust state-of-the-art ML solutions. While recruitment is still underway, we outline our approach to inclusion of underserved community in ML methods development. Through this

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
T. Zhang et al. (Eds.): Deep-Breath 2025, LNCS 16142, pp. 289–299, 2026.
https://doi.org/10.1007/978-3-032-05559-0_29

study, we aim to demonstrate the feasibility of POCUS for breast cancer screening in underserved African settings. The resulting dataset—the first of its kind—will facilitate ML-driven screening tools tailored for Black African women, supporting earlier detection and aligning with global targets to decrease breast cancer mortality and morbidity.

Keywords: Breast Cancer Screening · Breast imaging · POCUS · Africa · Deep Learning · Community-based participatory research

1 Introduction

Breast cancer is among the most common types of cancer worldwide [1]. Over the past 30 years, its incidence rate has increased globally due to compounding reproductive/hormonal risk profiles like advanced age at first birth, less breastfeeding, oral contraceptives, behavioral risk factors (including physical inactivity and consumption of processed foods), and improved cancer detection, among others [2]. While the incidence rate of breast cancer is reported to be higher in high income countries, the rapid economic transition in Sub Saharan Africa (SSA) is driving the exponential rise in the number of new cases in the region, faster than in high income countries [2, 3]. By 2040 or in 17 years to be exact, the incidence rate of breast cancer in SSA is projected to double [4].

Black African women have the highest breast cancer mortality rate [5, 6] and the youngest demographic of breast cancer patients (median age of 44 years on diagnosis) [1–8]. In sub-Saharan Africa, every second black woman diagnosed with breast cancer dies within five years [2, 4]. This starkly contrasts with the US, where only one in five black women and one in ten white women face similar outcomes [4]. The high mortality rates are largely attributed to late-stage diagnosis, where access to prompt and early detection are limited [7, 8]. Some studies have demonstrated that long travel distance to diagnostic facilities was associated with delay and late-stage diagnosis [9]. For SSA to achieve The World Health Organization (WHO) Global Breast Cancer Initiative (GBCI) target to reduce global breast cancer mortality by 25% by 2030 (in 7 years), innovative population-based interventions aimed at implementing the three pillars of WHO GBCI (health promotion for early detection, timely diagnosis, and comprehensive treatment) [5] must be developed.

Within sub-Saharan Africa, significant disparities exist, with black women in countries like Namibia and South Africa facing lower survival rates (56% and 59%, respectively) compared to their white counterparts (90% in Namibia) [2]. The situation is even direr in East and West Africa, where countries like Uganda and Nigeria report survival rates as low as 44–47% and 36%, respectively [2]. The need for early detection and treatment cannot be over-emphasized and concerns of shortage of skilled personnel, poor imaging infrastructure, huge patient burden, and insufficient healthcare financing are typically experiences in low-resource settings such as Africa. As low-cost imaging technologies such as Point of care Ultrasound (POCUS) show clinical promise [10], particularly hand-held devices, there is need to exploit their accessibility in improving

breast cancer screening through enhancement of lesion detection and characterization with machine learning (ML).

Here, we introduce the Precision Cancer Care in Africa (PRECISE) [11] framework for ML-guided community-based screening programs aimed at democratizing imaging capacity to increase access to high value diagnostic care for vulnerable populations in Africa and other underserved communities. Through PRECISE, we are curating an open-access dataset (Africa Breast Imaging Dataset for Equitable Cancer Care, ABreast Data) and organizing a data challenge (PRECISE-Africa Breast Ultrasound Segmentation and Classification Challenge, PACE) [11] to benchmark promising ML solutions for accessible breast cancer screening.

2 Materials and Method

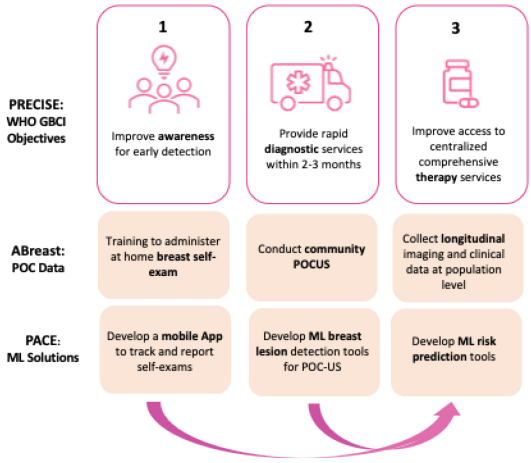


Fig. 1. The PRECISE framework to prevent breast cancer deaths in SSA*. *Figure is adapted from the IARC Evidence Summary Brief No. 1 and WHO Global Breast Cancer Initiative [5].

2.1 The Precision Cancer Care in Africa (PRECISE) Framework

The PRECISE framework (Fig. 1) integrates ML solutions arising from locally sourced clinical and imaging data, with community-based participatory research methods to create culturally aware screening initiatives for adoption of point-of-care (POC) imaging solutions, such as POCUS. The framework developed pathways towards use of ML-guided POC imaging as standard-of-care in underserved and inaccessible communities. The framework was formulated through the inaugural PRECISE Symposium [12] held on September 23 to 25th, 2024 at Abuja, Nigeria and funded by the University of Pennsylvania Penn Global Grant and Lacuna Fund on Sexual, Reproductive, and Maternal Health and Rights. This pivotal initiative was a collaborative effort between The University of

Pennsylvania Center for Global and Population Health Research in Radiology and the National Institute for Cancer Research and Treatment Nigeria (NICRAT), in partnership with the Project Pink Blue, Global Health Catalyst, Ernest Cook Ultrasound Research and Education Institute (ECUREI), Uganda, the Medical Artificial Intelligence Lab, Nigeria, Crestview Radiology Limited, Nigeria, and the Consortium for Advancement of MRI Education and Research in Africa (CAMERA).

The three-day hybrid event comprised of invited talks, panel discussions, interactive breakout sessions, hands-on training workshops in community engagement and POCUS imaging, an ML hackathon and networking opportunities. This brought together an interdisciplinary network of clinicians, scientists, industry partners, people with lived experiences, and community members to generate new insights into cancer diagnosis and treatment specific to the African context. A one-week hackathon was organized a week before the symposium from September 16th and concluded on day 1 of the symposium on September 23rd. The hackathon was aimed at developing regional ML imaging capacity through mentored training in ML model development, evaluation, and validation for lesion detection and classification on ultrasound data.

2.2 The African Breast (ABreast) Imaging Datasets

The PRECISE framework was immediately implemented, in a pilot prospective study (ABreast) to close gaps in breast cancer screening, by providing the first publicly accessible training and evaluation imaging datasets for ML applications in Black African women. We partnered with grassroot breast cancer advocacy groups (Project Pink Blue, Breast Without Spots, My Body My Asset), and people with lived experiences to conduct the ABreast study (Fig. 2). The study provides breast cancer awareness, at home breast self-exam training, destigmatization training, and POCUS-based screening (Lumify, Philips Healthcare) to underserved communities within 50 km of Lagos, Nigeria and Kampala, Uganda – largest cities in both countries and were access to imaging facilities for diagnostic validation can be reliably and safely performed. The ABreast study will identify breast lesions on POCUS breast scans of 100 consenting Ugandan and Nigerian women 18 years or older scanned in the community. The women will receive follow-up mammograms, breast magnetic resonance imaging (MRI), biopsy pathological data, where indicated, and comprehensive clinical profiling (demographics, medical history, relevant nutritional information, and blood samples for future genomic analysis), at the study site diagnostic imaging centers (Crestview Radiology, Nigeria, and ECUREI, Uganda).

Data collection for the 1-year study began in December 2024 after study ethics approval, with the first community engagement event in Lagos, Nigeria. At each event, information on breast cancer care pathway, breast self-exam, and breast screening are provided to participants. Those that agree to a breast screening with POCUS are recruited to the study and given the letter of information and an informed consent form is obtained. Trained medical personnel (radiology resident, medical students, medical doctors) or community health workers, then collect demographics and clinical history via structured interviews, measure vital signs (Microfield Analogue Scales with height bar and Omron M1 basic Automatic Upper Arm Blood Pressure Monitor), and breast ultrasound scans are performed by trained medical personnel using the Lumify handheld POCUS system.

Participants with abnormal POCUS scans will receive counseling on navigating the clinical care and referred to the study centers for follow-up investigations as outlined above. All women in the community are invited to participate in the study, except for breastfeeding mothers, women currently undergoing treatment for any breast pathology or have other forms of cancer or known breast pathologies.

The ABreast data will be preprocessed following standard ultrasound [13], mammograms, and MR image analysis steps. All images will be annotated by two trained radiologists, radiology residents, or fellows using Labelbox (https://labelbox.com). Consensus labelling and scoring will be used to generate the reference annotations for delineating the tumor margins and scoring the image BI-RADS. All curated and annotated datasets will be made publicly available, first through the upcoming PACE Challenge and archived in the National Institute of Health (NIH) The Cancer Imaging Archive (TCIA), representing to our knowledge the first publicly available POCUS data in the world.

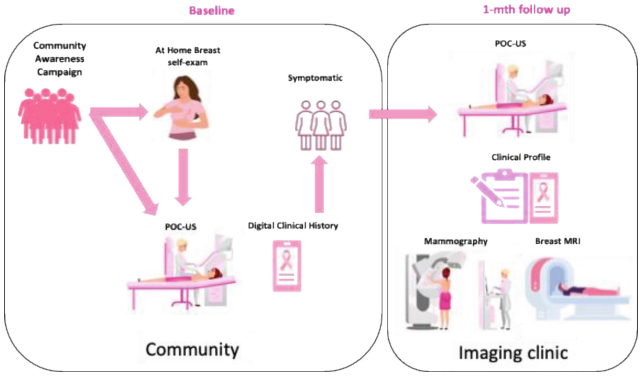


Fig. 2. Overview of the ABreast Study

2.3 PRECISE-Africa Breast Ultrasound Segmentation and Classification (PACE) Challenge

Since 60% of the ABreast datasets have been collected, the PACE challenge was initiated to use the current images to support development and evaluation of a single ML model proposed by participating teams. The model will be evaluated for its ability to perform multi-task learning, to simultaneously predict breast-level diagnostic category (no lesion, benign lesion, malignant lesion) – classification task and generate pixel-wise segmentation of any visible lesion – segmentation task.

The three-month challenge, started on June 20th, 2025, and runs in two phases: Phase 1, leverages retrospective breast conventional ultrasound images of 7354 patients (4574 benign, 2248 malignant, and 531 normal) from publicly available data (see challenge description on challenge webpage [11]) for ML training and validation. Phase 2 will introduce the unpublished prospectively collected ABreast POCUS dataset of 100 SSA females for model validation and test on hidden cases. The models submitted by participants will be evaluated using benchmarked expert annotations and ranked to determine

the best performing model for expanded validation in community settings, and in-depth exploration of the role of ML in POCUS. Performance will be assessed by averaging ranked scores across segmentation (normalized surface Dice, and lesion-wise dice similarity coefficient) and classification metrics (area under the curve, specificity, sensitivity at 90%, with equal weighting).

3 Results

3.1 The 2024 PRECISE Symposium, Hackathon and Framework

The framework developed through the PRECISE symposium aimed improved early detection and diagnosis of cancer in African populations. A total of 140 individuals registered to attend the 2024 PRECISE Symposium, 33 (~24%) were community health workers. Sixty-eight participants attended in person, while 22 actively attended virtually including speakers. A total 22 speakers gave invited presentations at 22 sessions and provided insights in two panel discussions (Think Tanks) over the three-day event (15 h). A parallel POCUS workshop with didactic lecture on principles of ultrasound imaging, hands-on POCUS training and demonstrations was provided over the three-day event to attendees with focus on its use for breast cancer screening to medical professionals and community health workers engaged in breast cancer awareness and screening. A snapshot of the program is shown in Fig. 3 and details of the three-day events are available on the 2024 symposium webpage [12].

Thirty participants from 7 African countries enrolled in the hackathon and were organized into four collaborative teams. Participants were provided with access to Google Collab Pro, to leverage High-Performance Computing (HPC) resources for the duration of the hackathon. The dataset consists of 780 high-quality ultrasound images from 600 female patients (25–75 years old), carefully selected from an initial 1,100 scans after preprocessing [14]. Each image, was captured using LOGIQ E9 and LOGIQ E9 Agile ultrasound systems, equipped with a 1–5 MHz ML6-15-D Matrix linear probe. These images represent scans from patients classified as either normal, benign, or malignant, and with lesion masks generated from expert segmentations. The dataset was split into 750 training and validation image sets and 30 held-out test sets. Each team submitted their model code and model performance was evaluated on the unseen test dataset, using two established metrics: Binary Cross-Entropy combined with Dice loss (BCE + Dice) and the Dice Similarity Coefficient (Dice). The performance of the proposed models is summarized in Table 1, in the order of ranked performance and based on higher Dice and BCE + Dice scores, showing case the feasibility of rapidly training and upskill researchers in region to build ML imaging solutions. The codes describing four proposed models are openly available on the hackathon repository [15].

Together, the knowledge synthesized from the PRECISE symposium are provided a robust framework to pilot data-driven cancer care studies in community settings (Fig. 3). The framework includes six overlapping themes, developed to: 1) address miseducation and stigma within the community, 2) enhance the use of medical imaging at the community to guide early detection, 3) deepen community education and screening through training and empowerment of community health workers (CHWs), 4) enhance active

patient engagement in research priority setting, conduct and dissemination, as collaborator with lived experience, 5) engage community with stakeholders including policy makers, researchers, clinical providers, and 6) implement population-specific artificial intelligence (AI) diagnostic methods.

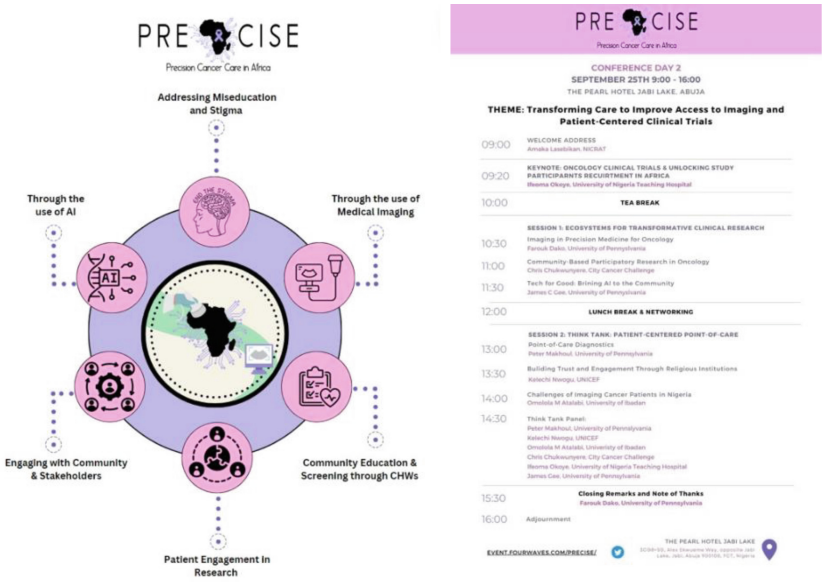


Fig. 3. The Precision Cancer Care in Africa (PRECISE) Framework.

Table 1. PRECISE-Hackathon 2024: Model Performance on Validation and Test Sets

Team	BCE + Dice	Dice score	Model
Team 4	27.4	0.76	Optuna Tuned Attention Unet
Team 2	3.2	0.40	Average Ensemble of U-Net and Attention Unet
Team 3	2.4	0.36	Attention Unet
Team 1	1.11	0.19	nnUNetv2

3.2 Abreast Study

To date, 18 unique community events have been conducted in townships within the twenty local government area in Lagos, Nigeria. From this a total of 474 women attended the events and were recruited, 426 consented to participate in the study (18 to 78 years old, 43 ± 5.88 years old) and 39 were found to have breast lesions on POCUS, representing approximately 8.6% of the study population. Of the Nigerian women with lesions, 14 (45%) have follow-up scans and clinical exams. A total of 46/359 women have never been pregnant, while the minimum age at first pregnancy among consented participants

was 15 years old, and the maximum age was 41 years old. Nineteen of the women reported a family history of breast issues, either benign or malignant, and 14 indicated a history of cancer in the family. None of the women have a history of smoking, while 36 (1%) reported consuming alcohol, 58% lead a sedentary lifestyle, and 26% have a body mass index (BMI) greater than or equal to 30.

In Uganda, a total of 91 community-dwelling women were recruited from communities that attend the Uganda Cancer Institute events. Of these, 23 had evidence of breast lesions on POCUS imaging. Across both sites and over 6 months of the study, 575 women have participated in community events, and 62 women (10.78%) have consented and had baseline scans. An illustration of the ABreast Study POCUS is shown in Fig. 4 of two randomly selected Nigerian cases, a scan in a normal female and another female with breast lesions and a BIRADS 2 score.

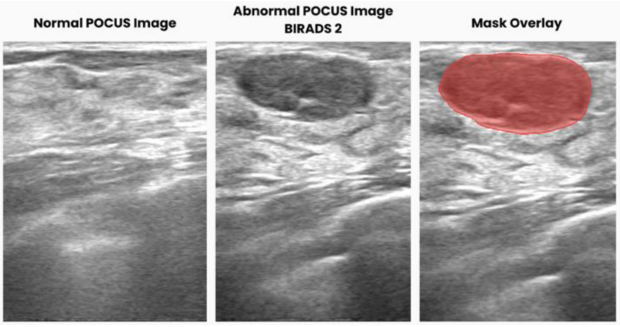


Fig. 4. Point-for-Care (POCUS) ultrasound images in two Nigerian females

3.3 PACE Challenge

Based on our capacity building efforts through the 2024 PRECISE symposium, particularly the hackathon and past experience organizing a population-specific imaging challenge [16], we anticipate 60 or more African researchers and clinicians will participate in the 2025 scaled up hackathon, running in 2025 as the PACE challenge. To date 23 teams have registered to participate in the challenge. The preprocessing and annotation of the ABreast data are underway with target to include 80 patient POCUS scans with breast lesions and 200 without lesions (normal scans) in the PACE challenge as test datasets.

4 Discussion

This work demonstrates the feasibility of using ML methods to guide POCUS for breast cancer screening tool in underserved sub-Saharan African communities. The detection of breast lesions in 10.9% of screened women emphasizes the presence and significant burden of breast abnormalities in this population and the possibility of increasing breast cancer incidence in SSA (2,3). The use of POCUS aligns with the WHO Global Breast Cancer Initiative, which advocates for innovative, population-based interventions such

as this accessible screening approach tailored to resource-limited settings [17]. Our proposed PRECISE framework which integrates ML in combination with community based participatory research and awareness, presents a new approach to minimize breast cancer trends, thorough enhanced point-of-care detection in the community.

While, ultrasound is particularly suited to the SSA demographic, particularly the predominance of large breast, it is equally affordable, accessible and portable compared to mammography. In the ABreast study, despite the apparent lesion detection with POCUS, only 45% of women with abnormal findings completed follow-up investigations, underscoring persistent barriers in the breast cancer care continuum. Reasons indicated for no follow-up include financial constraints (cannot take time off work), limited access to specialized care, transportation difficulties, and cultural stigma around cancer [4, 8]. Improving patient navigation, community education, and readily available healthcare infrastructure is essential to ensure timely diagnosis and treatment following screening. The study also identified lifestyle and reproductive factors which are relevant to breast cancer risk. High rates of sedentary behavior (58%) and obesity (26%) among participants are noted, given their association with increased breast cancer risk and poorer prognosis [2]. The reproductive profiles, including age at first pregnancy and nulliparity, further highlight the need for personalized risk assessment in screening programs. The absence of smoking and minimal alcohol consumption may reflect regional behavioral patterns that differ from Western populations, suggesting that prevention strategies should be culturally tailored.

A key innovation of this work is the development of a machine learning (ML)-ready dataset from clinical and imaging data, using the ABreast data to crowdsource benchmarked ML solutions through the PACE challenge. Existing ML models for breast lesion detection are predominantly trained on non-African populations. The tailored dataset will facilitate the creation of AI-driven diagnostic tools optimized for breast cancer screening in underserved communities (Black African women), but applicable to all settings, potentially improving diagnostic accuracy, reducing operator dependency, and expanding screening capacity in settings with limited radiological expertise [10]. Our work has important implications for breast cancer management in SSA. Community-based screening programs using POCUS, supported by ML interventions, could significantly enhance early detection. However, to translate screening into improved outcomes, health systems must address barriers to follow-up care and integrate lifestyle interventions targeting obesity and physical inactivity. Capacity building for healthcare workers in ultrasound use and local skills training in AI technology development and implementation will be critical for sustainable scale-up and wider clinical adoption [18].

While, the PRECISE framework presents a holistic approach to cancer screening, the relatively small sample size of the ABreast study (funding availability or lack of) and the high attrition (low follow-up) rates, could introduce selection bias, limit generalizability, and the impact of the proposed framework. We are working together with community partners, research and clinical collaborators to build stronger networks that can support longitudinal studies in larger cohorts, longer follow-ups, interventions to improve access and adherence to diagnostic and treatment pathways. This is the focus of the 2nd PRECISE Symposium to be held in September 1–3, 2025, in Nigeria.

In general, the PRECISE framework implemented through the ABreast study and PACE data Challenge demonstrates the promised of ML-supported POCUS imaging, as feasible and effective breast cancer screening tool in underserved populations. Coupled with community-based interventions and awareness intiatives, this approach holds promise for improving early detection and reducing breast cancer mortality among the most vulnerable - Black African women. Addressing systemic barriers and integrating comprehensive prevention strategies will be essential to realizing the full potential of these advances.

Acknowledgments. This project is partly supported by the Lacuna Fund on Sexual, Reproductive and Maternal Health and Rights (PIs: Kawooya and Anazodo) and the Penn Global Grant (PI: Tolulope Oyetunde). The authors acknowledge the invaluable contributions of the 2024 PRECISE Symposium speakers, attendees and partners (Breast without Spots, City Challenge, Network of People Impacted with Cancer in Nigeria, and the National Institute for Cancer Research and Treatment (NICRAT), Nigeria). We recognize and thank the patients with lived experiences who have participated in community events, lent their voices and contributed their data to this project.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Fitzmaurice, C., et al.: Global, regional, and national cancer incidence, mortality, years of life lost, years lived with disability, and disability-adjusted life-years for 32 cancer groups, 1990 to 2015: a systematic analysis for the global burden of disease study. JAMA Oncol. **3**, 524–548 (2017)
2. Adewale Adeoye, P.: Epidemiology of Breast Cancer in Sub-Saharan Africa. IntechOpen (2023)
3. 5772/intechopen.109361
4. Adeloye, D., et al.: Estimating the incidence of breast cancer in Africa: a systematic review and metaanalysis. J. Glob. Health **8**(1), 010419 (2018)
5. World Health Organization (WHO), International Agency for Research on Cancer (IARC): Breast Cancer Outcomes in Sub-Sharan Africa. IARC Evidence Summary Brief No. 1, March 2021. https://www.iarc.who.int/wp-content/uploads/2021/03/IARC_Evidence_Summary_Brief_1.pdf
6. Galukande, M., Wabinga, H., Mirembe, F.: Breast cancer survival experiences at a tertiary hospital in sub-Saharan Africa: a cohort study. World J. Surg. Oncol. **13**, 1–8 (2015)
7. Sankaranarayanan, R., et al.: Cancer survival in Africa, Asia, and Central America: a population-based study. Lancet Oncol. **11**, 165–173 (2010)
8. Gakwaya, A., et al.: Cancer of the breast: 5-year survival in a tertiary hospital in Uganda. Br. J. Cancer **99**, 63–67 (2008)
9. Togawa, K., et al.: Geospatial barriers to healthcare access for breast cancer diagnosis in sub-Saharan African settings: The African Breast Cancer-Disparities in outcomes cohort study. Int. J. Cancer **148**(9), 2212–2226 (2021)

10. Baloescu, C., Parhar, A., Liu, R., Wanjiku, G.W.: Effect of point-of-care ultrasound on clinical outcomes in low-resource settings: a systematic review. Ultrasound Med. Biol. **48**(9), 1711–1719 (2023)
11. https://precisesymposium.org
12. https://event.fourwaves.com/precise
13. Vallez, N., Bueno, G., Deniz, O., et al.: BUS-UCLM: Breast ultrasound lesion segmentation dataset. Sci. Data **12**(242) (2025)
14. Al-Dhabyani, W., Gomaa, M., Khaled, H., Fahmy, A.: Dataset of breast ultrasound images. Data Brief **21**(28), 104863 (2019). https://doi.org/10.1016/j.dib.2019.104863.PMID:31867417;PMCID:PMC6906728
15. https://github.com/CAMERA-MRI/PRECISE-Hackathon-2024
16. Adewole, M., et al.: The Brain Tumor Segmentation (BraTS) Challenge 2023: Glioma Segmentation in Sub-Saharan African Patient Population (BraTS-Africa). arXiv:2305.19369 (2023)
17. The WHO Global Breast Cancer Initiative (GBCI). https://www.who.int/initiatives/global-breast-cancerinitiative. Accessed 19 Jun 2023
18. Sood, R., et al.: Ultrasound for breast cancer detection globally: a systematic review and meta-analysis. J. Glob. Oncol. **5**(5), 1–17 (2019). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6733207/



Mammographic Image Generation Using Generative Cellular Automata

Lea Schwarz[1(✉)], Ricardo Montoya-del-Angel[1], Marawan Elbatel[2], and Robert Marti[1]

[1] Computer Vision and Robotics Institute, University of Girona, Girona, Spain
lea-schwarz@posteo.de
[2] The Hong Kong University of Science and Technology, Hong Kong SAR, China

Abstract. The application of generative models for the synthetic expansion of medical image datasets has been explored in recent research, offering new opportunities to combat data scarcity and improve computer-aided diagnosis (CAD) systems. In this study, we explore the use of a generative framework based on neural cellular automata (NCA) for the synthesis of mammographic images. While UNet and similar models focus on global coherence, NCAs inherently prioritize local dependencies through repeated, neighborhood-based updates. This feature facilitates capturing local details such as microcalcifications, subtle spiculations, and fine tissue textures, crucial for accurate diagnosis with mammographic images. To our knowledge, this is the first implementation of this type of framework for generating synthetic mammograms. The framework was adapted to process images from the VinDr-Mammo dataset, supporting both full mammograms and extracted findings in a highly configurable workflow. We integrated a custom breast density classifier to evaluate the impact of the addition of synthetic images on downstream classification performance. Qualitative and quantitative evaluations demonstrate that the adapted model is capable of generating perceptually realistic mammograms, with Kernel Inception Distance scores between 0.015 and 0.032. The addition of synthetic data to the density classifier's training dataset improved macro F1-scores by up to 12% in underrepresented density classes across multiple dataset configurations, although the effects on binary classification metrics remained mixed. These findings suggest that neural cellular diffusion models hold promise for mammographic image synthesis and further work is warranted to validate these results in clinical settings and with additional datasets.

Keywords: Mammographic Image Generation · Neural Cellular Automata · Computer Aided Diagnosis

1 Introduction

1.1 Mammography and Breast Density

Mammography has been the clinical standard for breast cancer screening for decades due to its accessibility and cost-effectiveness. Since the introduction of

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
T. Zhang et al. (Eds.): Deep-Breath 2025, LNCS 16142, pp. 300–309, 2026.
https://doi.org/10.1007/978-3-032-05559-0_30

full-field digital mammography (FFDM) in 2000 [3], digital imaging has enabled more flexible and reliable interpretation and data exchange compared to previously employed screen-film systems.

Despite these advances, reduced diagnostic sensitivity prevails in women with dense breast tissue, where fibroglandular structures can obscure lesions [2]. Breast density, categorized with four different density classes (A-D) via the BI-RADS scale [4], is both an independent risk factor for breast cancer as well as a factor complicating the diagnosis [8].

1.2 Synthetic Image Generation for Mammography

To address data scarcity and imbalanced datasets, generative models for mammographic image generation have been investigated. Previous work focused on Generative Adversarial Networks (GANs) to augment mammographic datasets [7]. Further, ROI-specific models, such as ROImammoGAN [11], as well as contextual GANs [12] were proposed for lesion-specific synthesis. But since GAN-based models often struggle with training instability and limited diversity in the generated samples, it is of interest to explore further architectures.

Diffusion models have recently emerged as a robust alternative, offering improved image fidelity and a high level of controllability. Conditional diffusion approaches, such as MAM-E [1], have the ability to synthesize full-field mammograms or perform lesion inpainting with controllable radiological attributes.

1.3 Neural Cellular Automata for Medical Image Generation

Neural Cellular Automata (NCA) are a class of generative models where a neural network regulates the evolution of each cell in a grid, which can lead to complex global structures emerging from local interactions [9].

NCA have recently been adapted for medical image synthesis tasks, with a notable application being the Generative Cellular Automata (GeCA) framework, which demonstrated competitive results in the synthesis of high-resolution retinal fundus and optical coherence tomography (OCT) images [5].

In GeCA, an image is modeled as a grid of "pix-cells", each entailing a local state vector consisting of a constant input value, a constant positional encoding, the predicted output pixel value, and a hidden state. The iterative evolution of these pix-cells is regulated by a localized Diffusion Transformer (DiT) block and happens based on a pix-cell's own state as well as the state of neighboring pix-cells. Further, the model also integrates biologically inspired mechanisms such as fusion and mitosis, which perform spatial downscaling and upscaling during the generative process.

Despite producing promising results, the framework's application to mammographic image synthesis has not yet been explored. The present study investigates the potential of adapting the GeCA framework to generate realistic mammograms on a highly flexible data setup.

Beyond adapting the framework, this study introduces a highly configurable workflow capable of generating both full mammograms and extracted findings,

using the VinDr-Mammo dataset as a test case. Furthermore, a custom breast density classifier is integrated to evaluate the impact of synthetic image augmentation on downstream classification performance.

2 Methodology

2.1 Data and Annotation Setup

For training and testing purposes, we used the VinDr-Mammo dataset, a publicly available collection of full-field digital mammograms with breast-level radiological annotations [10].

A customized CSV annotation format was established to flexibly select and consistently manage data across training, validation, and testing splits. This file includes image identifiers, split information, labels (either as categorical or multi-label lists), and optional fold indicators for cross-validation setups. All images are preprocessed (cropping, resizing to 128×128 pixels, and vendor text removal) and stored in PNG format.

To further standardize data handling throughout the pipeline, we developed MammoDataset, a subclass of the torch Dataset class. It loads images and their corresponding labels based on the customized annotation CSV, supporting both one-hot encoded multi-label annotations and single-label formats, depending on the column chosen to label the images for training and generation.

2.2 Adapting the Diffusion Framework for Mammography

To synthesize mammograms, significant adaptations were made to the original neural cellular automata-based framework [5]. Figure 1 depicts the schematic process traversed in the framework. The full implementation can be found on GitHub https://github.com/VICTORIA-project/GeCA_mammography.

Input images are encoded into latent representations using a pretrained Variational Autoencoder (VAE) and cached for training. The NCA model treats each image as a grid of cells and iteratively updates their states using localized Diffusion Transformer blocks. Training optimizes the model to reconstruct real images from noise, using classifier-free guidance and stratified cross-validation. After training, the framework generates new mammograms by stochastically evolving cell states from a single initialized cell to fill the whole grid. The sampling script allows label-specific generation and selected numbers of images per label to facilitate targeted class-balancing. Finally, assessments of the quality, diversity, and generalization performance of the generated images are performed, employing the Kernel Inception Distance (KID), Learned Perceptual Image Patch Similarity (LPIPS), and the Generalization Gap (GG) as metrics, as well as a downstream breast density classification task.

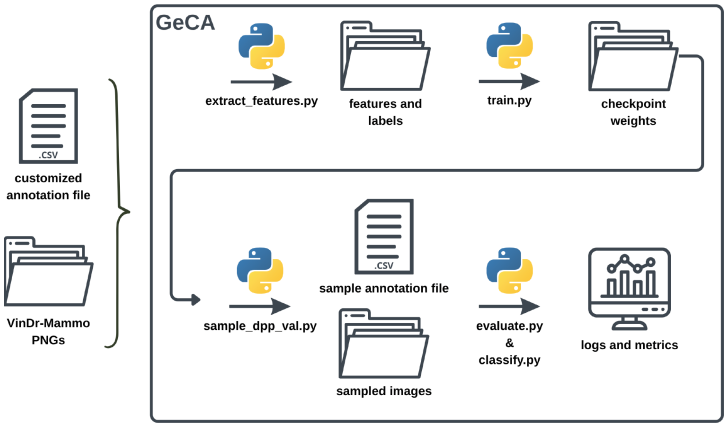


Fig. 1. A schematic representation of the process followed in the pipeline.

2.3 Downstream Evaluation via Breast Density Classification

To assess the clinical utility of the generated synthetic images, we implemented a breast density classification pipeline.

To fulfill the task, a simple Convolutional Neural Network (CNN) was implemented, consisting of:

- Two convolutional layers (ReLU activation + max pooling)
- Two fully connected layers
- CrossEntropyLoss for optimization
- Adam optimizer (learning rate: 0.001)
- Early stopping based on validation macro F1-score

For the purpose of comparing the performance of the classifier when trained on different datasets - with and without synthetic expansion in this case - multiple evaluation metrics were logged, such as the Macro-averaged F1-score, confusion matrices and per-class precision, recall, F1-score and support.

In addition to the classification of all four density classes, labels were grouped into a clinically meaningful binary split: A+B (low density) vs. C+D (high density), allowing for the calculation of corresponding binary classification metrics.

The classifier expects the same annotation CSV format as the rest of the framework. To facilitate seamless classifier training on mixed datasets containing both real as well as synthetic images, a script generating appropriate files has been implemented.

2.4 Experiments

To systematically evaluate the framework and its strengths and weaknesses, multiple runs were performed across different data subsets and label configurations,

as detailed in Table 1. The framework enables flexible selection of filtering conditions and training labels, in the tried configurations, filtering was performed either on mammogram acquisition machine models (Mammomat Inspiration, Planmed Nuance or both) or in one case only images containing masses according to VinDr-Mammo annotations are used. Training labels are selected to be taken from an annotation file column, in the investigated cases either finding categories, breast density or in one case the mammograph model.

Due to limited computational resources, most trainings were conducted at an image resolution of 128×128 pixels. Aside from this one exception, training settings were held constant across the runs: batch size was fixed to 32, training conducted on two GPUs, specifically an NVIDIA A40 GPU with 46GB memory and an NVIDIA A30 GPU with 24GB memory, and a validation was performed every 50 epochs, with checkpoints saved when the Kernel Inception Distance (KID) improved.

The number of training epochs depended on dataset size: runs with over 12,000 images trained for 1,000 epochs, while smaller datasets (<3,500 images) were trained for 2,500 epochs. Table 2 summarizes the dataset split sizes and training durations.

Table 1. Data configurations used for model training.

Run Identifier	Filtering Conditions	Label Column
MA_PL_FI	Mammomat Inspiration or Planmed Nuance	Finding categories
MA_PL_DE	Mammomat Inspiration or Planmed Nuance	Breast density
MA_FI	Mammomat Inspiration	Finding categories
MA_DE	Mammomat Inspiration	Breast density
PL_FI	Planmed Nuance	Finding categories
PL_DE	Planmed Nuance	Breast density
MAS_MO	Finding categories: Mass	Model

Breast Density Classifier Experiments. For the breast-density configurations (MA_PL_DE, MA_DE, PL_DE) including the higher-resolution one, we ran classification experiments to evaluate how targeted synthetic augmentation affects performance. To correct the severe class imbalance, equal numbers of synthetic images were generated for the underrepresented density categories (A, B, and D): 10,000 per class for MA_PL_DE and MA_DE, and 2,000 per class for PL_DE. By comparing classifier results on these augmented datasets against the baseline trained solely on real images, a systematic assessment of the potential of synthetic augmentation to boost downstream classification accuracy was facilitated.

Table 2. Dataset split sizes and training epochs.

Run Identifier	Training	Test	Validation	Epochs
MA_PL_FI	15,994	4,725	799	1,000
MA_PL_DE	16,093	4,744	870	1,000
MA_FI	12,716	3,786	611	1,000
MA_DE	12,649	3,801	674	1,000
PL_FI	3,220	934	214	2,500
PL_DE	3,335	935	163	2,500
MAS_MO	1,516	268	48	2,500

3 Results and Discussion

3.1 GeCA Training Outcomes

After adapting the Generative Cellular Automata (GeCA) framework to function on mammographic images, eight training runs with varying data configurations were executed. Metrics collected during training can be viewed in Fig. 2.

The training loss consistently demonstrated rapid convergence across all runs, indicating effective optimization. Notably, the MAS_MO configuration, containing the smallest number of images, achieved steeper loss declines. This likely stems from reduced image diversity, enabling faster learning of characteristics.

The Kernel Inception Distance (KID) mean values followed the expected pattern of a sharp early decline followed by stabilization, suggesting early developments towards realistic synthetic images. Interestingly, configurations using finding categories as labels generally reached lower KID mean values than their breast density counterparts, likely due to the localized nature of findings versus the more global patterns of breast density, especially given the low image resolution in most runs.

3.2 Synthetic Image Quality Evaluation

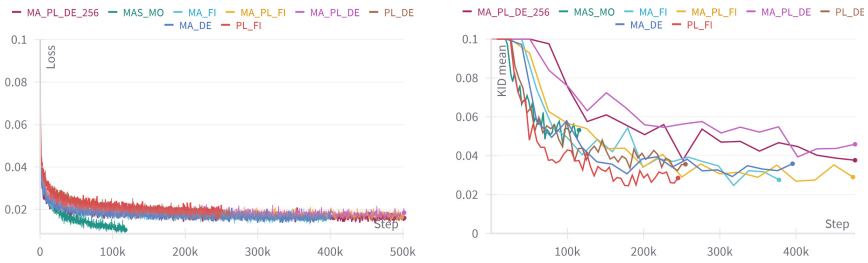


Fig. 2. Training loss and KID mean throughout GeCA trainings on mammographic images.

Qualitative Assessment. Visual inspection of synthetic images proved that GeCA effectively generated mammograms showing discernible background, breast boundary, and tissue structures. Real and synthetic image examples for all four breast density categories can be seen in Fig. 3.

In multi-device configurations, the synthetic images reflected distinctions between acquisition machines in pixel value patterns, indicating the model's capacity to capture and reproduce dataset diversity. Still, underrepresented classes showed limitations in perceptual clarity.

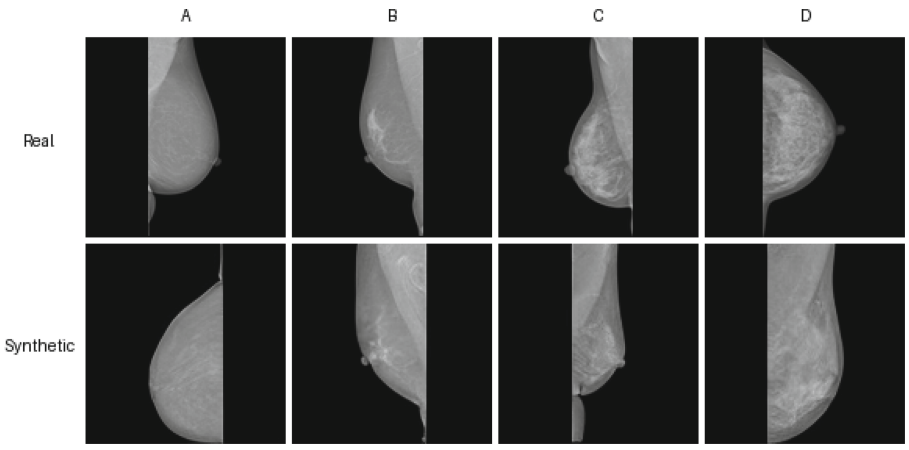


Fig. 3. An exemplary grid of real images for each density category in the upper row and synthesized images for the same classes in the bottom row, taken from the data configuration MA_PL_DE. The labels A, B, C, and D refer to the respective breast density, where A denotes the least and D the most dense tissue.

Quantitative Metrics. The quantitative evaluation of synthesized mammograms using KID, Learned Perceptual Image Patch Similarity (LPIPS), and Generalization Gap (GG) metrics confirmed these trends, though they produced comparably better results in the larger datasets than the ones with fewer images.

Compared to the benchmarks reported by Elbatel et al. [5], the mammogram KID means (0.0150.032) were comparable to or better than those for OCT images (0.0491) but did not reach Fundus image results (0.00742). LPIPS means (0.260.35) were lower than those reported for fundus (0.39) and OCT (0.53) images, denoting a slightly reduced diversity in the resulting synthetic datasets, which could be caused by the low image resolution.

The GG metric demonstrated robust alignment of the synthetic images with the distributions in the underlying real data, especially in larger datasets. Settings MA_PL_FI and MA_PL_DE achieved values close to zero, while for the case of MAS_MO a higher GG value was reached, indicated possible overfitting, likely due to low dataset size and diversity [6].

3.3 Breast Density Classification Performance

To evaluate the practical value of synthetic images for a downstream application, breast density classification tasks were performed on three low-resolution configurations (MA_PL_DE, MA_DE, PL_DE). We compared the classifier's performance when trained on exclusively real datasets, and targeted augmentation (expanding underrepresented classes).

Macro F1-Score Results Targeted augmentation was able to consistently improve macro F1-scores in all configurations and boost class-wise F1-scores, as shown in Table 3, particularly for the underrepresented classes (e.g. Density D).

Table 3. The macro F1-scores obtained for each configuration with original images compared to targeted augmentation.

	MA_PL_DE	MA_DE	PL_DE
Original dataset	0.429	0.432	0.418
Augmented dataset	0.464	**0.552**	0.435

Binary Classification Metrics. In the case of binary classification, the outcomes varied considerably depending on the configuration. In MA_PL_DE, targeted augmentation improved accuracy (0.936 to 0.952), F1-score (0.966 to 0.975), and recall (0.976 to 0.995). On the other hand, MA_DE and PL_DE experienced declines across all metrics, as seen in Table 4. This inconsistency may be caused by the dataset size: larger datasets produced higher-quality synthetic samples, whereas smaller ones lacked sufficient diversity for effective augmentation. This also obeys the usual requirements of diffusion-based models, which achieve higher performance when the dataset size increases. In our case, the good performance observed in all three configurations with real data can only be further enhanced when combining both Mammomat Inspiration or Planmed Nuance images.

Table 4. The binary F1-scores obtained for each configuration with original images compared to targeted augmentation.

	MA_PL_DE	MA_DE	PL_DE
Original dataset	0.966	0.961	0.95
Augmented dataset	**0.975**	0.957	0.933

3.4 Conclusions

Our implementation of the GeCA framework for mammographic imaging demonstrated robust performance in synthesizing low-resolution mammograms, achieving more consistent image quality when conditioning the model with finding categories rather than breast density labels. Higher synthetic quality is attained in larger and more diverse datasets. Synthetic image augmentation proved especially valuable for improving classification of underrepresented classes, provided sufficient dataset diversity was available.

The primary limitation of this study was the restricted image resolution (128×128px). Nevertheless, the promising results in both image synthesis quality and downstream classification performance encourage future exploration of higher-resolution image generation. Further research should encompass evaluation on additional mammographic datasets and the targeted synthesis of lesion-only images, a feature already supported by our current pipeline but not yet explored. Finally, assessing the clinical relevance of generated images within validated CAD workflows could provide important insight into their real-world utility.

Acknowledgments. This research was possible thanks to funding from the VICTORIA project (PID2021-123390OB-C21) from the Ministerio de Ciencia, Innovación y Universidades of Spain.

Conflicts of Interests/Competing InterestsDisclosure of Interests.. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Montoya-del Angel, R., Sam-Millan, K., Vilanova, J.C., Martí, R.: MAM-E: Mammographic synthetic image generation with diffusion models. Sensors **24**(7), 2076 (2024). https://doi.org/10.3390/s24072076, http://arxiv.org/abs/2311.09822, arXiv:2311.09822 [eess]
2. Aristokli, N., Polycarpou, I., Themistocleous, S.C., Sophocleous, D., Mamais, I.: Comparison of the diagnostic performance of Magnetic Resonance Imaging (MRI), ultrasound and mammography for detection of breast cancer based on tumor type, breast density and patient's history: a review. Radiography **28**(3), 848–856 (2022)
3. Conant, E.F., Maidment, A.D.: Update on digital mammography. Breast Dis. **13**(1), 109–124 (2001). https://doi.org/10.3233/BD-2001-13113, https://journals.sagepub.com/doi/full/10.3233/BD-2001-13113
4. D'Orsi, C.J., Sickles, E.A., Mendelson, E.B., Morris, E.A.: 2013 ACR BI-RADS Atlas: breast imaging reporting and data system. American College of Radiology (2014). google-Books-ID: nhWSjwEACAAJ
5. Elbatel, M., Kamnitsas, K., Li, X.: An organism starts with a single pix-cell: a neural cellular diffusion for high-resolution image synthesis (2024). http://arxiv.org/abs/2407.03018, arXiv:2407.03018 [cs]

6. Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., Aila, T.: Training generative adversarial networks with limited data (2020). https://doi.org/10.48550/arXiv.2006.06676, http://arxiv.org/abs/2006.06676, arXiv:2006.06676 [cs]
7. Korkinof, D., Rijken, T., O'Neill, M., Yearsley, J., Harvey, H., Glocker, B.: High-resolution mammogram synthesis using progressive generative adversarial networks (2019). https://doi.org/10.48550/arXiv.1807.03401, http://arxiv.org/abs/1807.03401, arXiv:1807.03401 [cs]
8. McCormack, V.A., dos Santos Silva, I.: Breast density and parenchymal patterns as markers of breast cancer risk: a meta-analysis. Cancer Epid. Biomark. Prev. **15**(6), 1159–1169 (2006). https://doi.org/10.1158/1055-9965.EPI-06-0034, https://doi.org/10.1158/1055-9965.EPI-06-0034
9. Mordvintsev, A., Randazzo, E., Niklasson, E., Levin, M.: Growing neural cellular automata. Distill **5**(2), e23 (2020). https://doi.org/10.23915/distill.00023, https://distill.pub/2020/growing-ca
10. Nguyen, H.T., et al.: VinDr-Mammo: a large-scale benchmark dataset for computer-aided diagnosis in full-field digital mammography (2022). https://doi.org/10.1101/2022.03.07.22272009, https://www.medrxiv.org/content/10.1101/2022.03.07.22272009v1, pages: 2022.03.07.22272009
11. Oyelade, O.N., Ezugwu, A.E., Almutairi, M.S., Saha, A.K., Abualigah, L., Chiroma, H.: A generative adversarial network for synthetization of regions of interest based on digital mammograms. Sci. Rep. **12**(1), 6166 (2022). https://doi.org/10.1038/s41598-022-09929-9
12. Wu, E., Wu, K., Lotter, W.: Synthesizing lesions using contextual GANs improves breast cancer classification on mammograms (2020). https://doi.org/10.48550/arXiv.2006.00086, http://arxiv.org/abs/2006.00086, arXiv:2006.00086 [eess]

On Tackling Domain Shift in Breast MRI Using Only Publicly-Available Data: Reproducible Breast Cancer Segmentation and pCR Prediction

Jessica Kächele[1,2,3(✉)], Dimitrios Bounias[1,2], Alexandra Ertl[1,2], and Klaus Maier-Hein[1,2,3,4,5,6]

[1] Division of Medical Image Computing, German Cancer Research Center (DKFZ), Heidelberg, Germany
jessica.kaechele@dkfz-heidelberg.de
[2] Medical Faculty Heidelberg, Heidelberg University, Heidelberg, Germany
[3] German Cancer Consortium (DKTK), DKFZ, Core Center Heidelberg, Heidelberg, Germany
[4] Pattern Analysis and Learning Group, Department of Radiation Oncology, Heidelberg, Germany
[5] National Center for Tumor Diseases (NCT), Heidelberg University Hospital (UKHD) and German Cancer Research Center (DKFZ), Heidelberg, Germany
[6] Faculty of Mathematics and Computer Science, Heidelberg University, Heidelberg, Germany

Abstract. Breast magnetic resonance imaging (MRI) offers high diagnostic capability for breast cancer, but comparison of machine learning (ML) methods has historically been limited by the lack of annotated public data. The MAMA-MIA challenge addresses this by providing voxel-level tumor segmentations for n = 1506 pre-treatment DCE-MRI cases. In this work, we present publicly reproducible ML models for two breast cancer-related tasks: breast tumor segmentation and prediction of pathological complete response (pCR) following neoadjuvant chemotherapy. For segmentation, we use a self-supervised pretraining strategy on n = 4799 public DCE-MRI volumes combined with a residual encoder architecture, achieving substantial performance above the baseline and strong fairness across demographic subgroups. For pCR classification, radiomics features extracted from segmented lesions are used with a calibrated gradient boosting classifier, emphasizing fairness and domain generalization. Our models demonstrate that high-performance and fair breast MRI analysis can be achieved using only public data, showing the potential for equitable and transparent ML oncological models.

Keywords: Breast cancer · Machine learning · Transfer learning

J. Kächele and D. Bounias—Equal contribution.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
T. Zhang et al. (Eds.): Deep-Breath 2025, LNCS 16142, pp. 310–319, 2026.
https://doi.org/10.1007/978-3-032-05559-0_31

1 Introduction

Breast cancer is the most common cause of cancer-related deaths among women globally [1]. The full multiparametric breast magnetic resonance imaging (MRI) protocol, which includes dynamic contrast-enhanced (DCE) acquisitions, offers the highest achievable imaging performance for breast cancer diagnosis by utilizing gadolinium-based contrast agent (GBCA) administration. Given the clinical value of breast MRI, numerous studies have explored applying machine learning (ML) to it using private data [2], but method development on public data has long been limited.

The primary reason behind this has been the lack of annotations available, especially in the form of voxel-level masks, despite the relatively high availability of public breast MRI images. The MAMA-MIA challenge [3] attempts to alleviate that, by providing $n = 1506$ annotations corrected/verified by 16 expert radiologists (averaging 9 years of experience) for previously publicly-available data, making method development using multi-institutional data accessible and enabling objective comparison of different strategies.

The reality of breast cancer diagnosis and treatment can differ between populations, scanners, acquisition protocols, and patient characteristics. Differences in scanners and acquisition protocols present domain shift-related methodological challenges, for which the MAMA-MIA challenge addresses by providing annotations for data sourced from different institutions with varying protocols and by having data from separate institutions for the test set. Further, the concept of fairness takes mainstage in the challenge, aiming to ensure performance is equal across differences in patient characteristics, which in the context of personalized medicine [4] is highly important, as it ensures attempts to tailor treatment to individual patients are not hindered by bad upstream performance.

In this work, we present our methods for breast cancer segmentation in breast MRI and pathological complete response (pCR) prediction based solely on the imaging data, with a focus on addressing domain shift and increasing fairness, as submitted to the MAMA-MIA challenge. The main contributions lie in using pre-training based on most unannotated breast MRI data combined with a large residual encoder for the segmentation task and using these annotations for feature-based pCR prediction with unsupervised label-shift calibration.

2 Materials and Methods

2.1 Data

Cohort Composition. The training dataset for this study was provided through the MAMA-MIA Challenge [3] and consists of $n = 1506$ pre-treatment DCE-MRI examinations sourced from four public collections of the cancer imaging archive (TCIA): I-SPY1 (n = 171), I-SPY2 (n = 980), Duke-Breast-Cancer-MRI (n = 291), and NACT-Pilot (n = 64). Each case includes one pre-contrast acquisition and up to five post-contrast phases, with both bilateral (30.9 %) and unilateral breast MRIs represented. The independent test set comprises 574 cases collected from European centers.

Training Annotations and Metainformation. As part of the challenge training data, tumor segmentations were provided (generated by first training and applying nnU-Net [5] to produce preliminary masks, which were then reviewed and corrected by 16 clinical experts to yield consensus 3D annotations). Clinical variables were available for over 90 % of patients and include breast cancer diagnosis, disease multifocality, molecular subtype, and pathological complete response (pCR) following neoadjuvant chemotherapy (NAC). Additional endpoints, available in subsets of the cohort, comprise overall survival status in more than 450 cases, detailed receptor profiles, time to recurrence or metastasis, NAC agents administered, post-treatment mastectomy rates, breast density, and other relevant clinical data. Additionally, region-of-interest (ROI) bounding boxes delineating the breast that contain the primary tumor are provided for all cases to guide analysis.

Cohort Comparison. A comparison between training and test cohorts highlights key differences in data acquisition and patient demographics. The training cohort included both axial (84.4 %) and sagittal (15.6%) scans, predominantly at 1.5 T (72.1 %) from GE (64.1 %), Siemens (27.3%), and Philips (8.6%) scanners, whereas the test cohort was nearly all axial with a more even field-strength split and different vendor proportions. Demographically, the test cohort had a higher fraction of patients over 70 years of age (18.5% vs. 2.1%) and postmenopausal status (65.3% vs. 27.9%) compared to the training set. Imaging parameters such as slice thickness (56.1 % of training scans at 2.00–2.49 mm vs. 40.4 % of test scans at 1.50–1.99 mm), pixel spacing, echo time, and repetition time also vary between cohorts. These disparities underscore the challenges of achieving robust model generalization across heterogeneous data sources.

2.2 Acquisition Selection

From the DCE-MRI data, we only utilized the three first T1-weighted acquisitions (i.e. the pre-contrast acquisition and the two initial post-contrast acquisitions), discarding the rest. This decision was because there was a variable number of acquisitions available for each examination, with two post-contrast acquisitions being the minimum that was always available. We decided to (i) not exclude examinations with a low number of acquisitions or (ii) try to generate the missing acquisitions or supplement them in some other way, as there are potential drawbacks to both approaches and performance could be adequate using data from early phases, given the increasing relevance of abbreviated protocols [6]. This also further increases applicability to said abbreviated GBCA-enhanced protocols (<10 min acquisition time, which as far as verifiable these time points fall within), which can be seen as a benefit of the method.

2.3 Task 1: Segmentation

Our segmentation approach builds on nnU-Net by employing the ResEnc architecture [7] and incorporating self-supervised pretraining on multiple public breast MRI datasets.

Table 1. Pretraining Datasets for Segmentation (Task 1). Time point 0 refers to the pre-contrast T1-weighted acquisition, while time points 1 and 2 pertain to post-GBCA T1-weighted acquisitions. * = Only examinations of these datasets that were not also part of the MAMA-MIA dataset were included.

Dataset	Exams (n)	Images (n)	Time points
Advanced-MRI-Breast-Lesions [14,15]	632	1,896	0,1,2
Duke-Breast-Cancer-MRI* [14,16]	624	1,872	0,1,2
ISPY1* [14,17]	42	42	1
EA1141 trial [14,18]	903	903	1
ACRIN-Contralateral-Breast-MR [14,19]	86	86	1
Total	**2,287**	**4,799**	

Pretraining. Table 1 lists the datasets used in this study. All volumes were first resampled to an isotropic voxel size of $1 \times 1 \times 1\,\mathrm{mm}^3$ using third-order spline interpolation. We then pretrained a 3D ResEncL architecture as a self-supervised masked autoencoder [8] built on the nnU-Net framework [9]. Pretraining was carried out for 1000 epochs with a batch size of 2.

Preprocessing. Volumes were first resampled to an isotropic voxel grid of $1\times1\times1\,\mathrm{mm}^3$ using third-order spline interpolation. Intensity values in each volume were then z-score normalized.

Fine-tuning. We used the nnU-Net framework with the ResEncL architecture [7], initializing all weights from our self-supervised pretraining (see Sect. 2.3). To accommodate the multi-channel MRI inputs, the stem weights were replicated. Fine-tuning was carried out using five-fold cross-validation on the MAMA-MIA training set.

Optimization began with an initial learning rate of 1×10^{-3}. We first warmed up the segmentation heads for ten epochs before including all network parameters in a second fifty-epoch warm-up phase. After these stages, the learning rate decayed according to a cosine-annealing schedule over the remaining 9940 epochs. During training, we applied the standard nnU-Net augmentation pipeline with elastic deformations, random intensity shifts, and spatial cropping.

Postprocessing. During inference, we averaged the probability maps produced by each member of the ensemble to obtain the final segmentation. All connected components falling outside the given ROI bounding box were then removed, eliminating spurious detections and restricting the segmentation to the anatomically relevant breast region.

Alternative Approaches Explored. Several variant models were evaluated to assess their impact on segmentation performance that didn't lead to a better performance:

- **Multitalent pretraining:** Pretraining using extensive and diverse collection of 3D medical images, combining various public datasets in a manner inspired by Multitalent [10].
- **Contrast-based pretraining:** Pretraining only on datasets containing a pre-contrast and two post-contrast phases to match the MAMA-MIA dataset.
- **Distance-based postprocessing:** Discarding all small connected components whose centroid lies farther than 50mm from the largest connected component.
- **Largest component retention postprocessing:** Retaining only the single largest connected component per case to enforce one contiguous lesion mask.

Evaluation. Segmentation performance was assessed using the Challenge-defined metrics: Dice similarity coefficient (DSC), normalized 95th-percentile Hausdorff distance (NormHD), and the fairness score. The DSC and NormHD are combined into a performance score:

$$\text{Performance Score} = \frac{1}{2}(\text{DSC} + (1 - \text{NormHD})) \tag{1}$$

Fairness is evaluated across three patient attributes, the age group, menopausal status, and breast density, by computing the performance disparity D_v between subgroups for each attribute v. The overall fairness score is then

$$\text{Fairness Score} = \frac{1}{|V|} \sum_{v \in V} (1 - D_v), \tag{2}$$

where $|V| = 3$. The final leaderboard ranking is determined by a weighted combination of performance and fairness:

$$\text{Total Score} = (1 - \alpha)\,\text{Performance Score} + \alpha\,\text{Fairness Score}, \tag{3}$$

where $\alpha = 0.5$.

2.4 Task 2: Classification

Pretraining and Segmentation. The same pretraining and segmentation methodology as the one used in task 1 applies, as the segmentation produced in task 1 were used as inputs in task 2.

Preprocessing. For the imaging data, we performed image intensity normalization to [0, 1] (by subtracting the minimum intensity and dividing by the range), as well as resampling to $0.7\times0.7\times2.0\,\mathrm{mm}^3$. For the segmentations, only the resampling was performed (using nearest neighbor interpolation).

Feature Extraction. From each of the three aligned DCE volumes of the training dataset cropped to the areas denoted by the segmentations produced for task 1, we extracted 321 radiomics features [11] per examination (i.e. 107/image). We removed any features that were constant across the cohort, leaving 195 features. For inference, we calculated these 195 features regardless of variance.

We used pyradiomics [11] for feature extraction (extracted full set of radiomics features; `binWidth = 25`, `sigma = [1,2,3]`, `enableCExtensions= True`, and `extractor.enableAllFeatures()`.

Feature Normalization. We then normalized each feature independently to [0, 1] (by subtracting the minimum intensity and dividing by the range). We stored the minimum and maximum values of each feature as computed on the training set. These statistics are then used to normalize incoming samples at inference.

Model. We trained a Gradient Boosting ensemble [12] of decision trees on 195 radiomic features to predict pathological complete response (pCR). An exhaustive grid search over

$$\text{max_depth} \in \{2, 3, 4\}, \qquad \text{max_features} \in \{\sqrt{p},\ \log_2 p,\ \text{all}\},$$

was carried out with 5-fold cross-validation on the training set, using the official Challenge metric to select the best configuration. All remaining hyperparameters were kept at their scikit-learn defaults, except for the learning rate (`0.10`), number of estimators (<`400`), subsample ratio (`1.0`), and loss function (`log_loss`). The best configuration ($\max_depth = 4$, $\max_features = \sqrt{p}$) was retrained on the full training cohort. We thresholded predicted probabilities at 0.5 to obtain binary labels. To correct for domain shift, we applied an unsupervised label-shift calibration targeting a 20% pCR rate, in line with the 19.8% median reported in a recent meta-analysis [13].

Alternative Approaches Explored. Prior to settling on our Gradient Boosting classifier, we evaluated a range of alternative algorithms and preprocessing strategies. We tested support vector machines with linear, RBF, sigmoid, and polynomial kernels; random forests with 100–600 trees, varying depths, and different minimum samples per leaf; and neural networks, trained with both standard cross-entropy and custom challenge-inspired loss functions. On the data side, we explored synthetic minority oversampling techniqu (SMOTE) to rebalance fairness-related subgroups and principal component analysis (PCA)

to reduce the features to 50 or 150 principal components, with performance increases been sparse across methodologies. We also experimented with ensembles of these diverse models, yet any marginal improvements failed to translate under domain shift on validation.

Evaluation. Classification performance was assessed using the Challenge-defined metrics: balanced accuracy and the fairness score. Fairness is measured across three patient attributes, the age group, menopausal status, and breast density, by comparing the true positive rates (TPR) of the most- and least-favored subgroups. For each attribute i,

$$\text{Fairness}_i = 1 - |\text{TPR}_{\text{group max}} - \text{TPR}_{\text{group min}}|. \tag{4}$$

The overall fairness score is the average over the three attributes:

$$\text{Fairness Score} = \frac{1}{3}\sum_{i=1}^{3} \text{Fairness}_i. \tag{5}$$

The final leaderboard ranking combines balanced accuracy and fairness:

$$\text{Total Score} = (1-\alpha)\,\text{Balanced Accuracy} + \alpha\,\text{Fairness Score}, \tag{6}$$

with $\alpha = 0.5$.

3 Results

3.1 Preliminary Results

The preliminary segmentation results on fold 0 reveal a steady improvement in segmentation accuracy across our model variants. The baseline nnUNet achieved a Dice score of 0.764, which increased to 0.798 with the ResEncUNetL backbone. Pretraining on datasets that include all three contrast phases (pre-contrast, first post-contrast, and second post-contrast) further increased the Dice to 0.808. Expanding the pretraining set to incorporate additional single-timepoint data yielded a Dice of 0.813, highlighting the benefit of greater data volume.

Our classifier achieved a balanced accuracy of 0.57 and a fairness score of 0.37, resulting in an overall score of 0.47.

3.2 Validation Set Results

On the validation set, our segmentation model achieved a DSC of 0.70 and a NormHD of 0.1. With a fairness score of 0.91, this resulted in a combined leaderboard score of 0.85, securing first place among all segmentation submissions at the time of writing.

3.3 Test Set Results

On the held-out test set, our segmentation model attained a DSC of 0.735 and a NormHD of 0.099. It achieved a performance score of 0.819, a fairness score of 0.96, and an overall score of 0.89. The classification model attained a balanced accuracy of 0.513 and a fairness score of approximately 0.70, yielding a total score of approximately 0.60[1].

4 Discussion

In this work, we present reproducible ML models for breast cancer segmentation and pCR prediction, trained solely on publicly-available data.

Our segmentation model demonstrates strong generalization capability across a multi-vendor DCE-MRI cohort and the independent test set, across nearly all demographic groups, illustrating the power of large-scale self-supervised pre-training.

Our pCR prediction model achieves a substantial balance between balanced accuracy and fairness, with a focus on attempting to mitigate domain shift concerns, which can be very useful given the diversity of scanners, field strengths, and acquisition protocols in breast MRI. While many approaches were found to provide high balanced accuracy results (up to 0.65) in our preliminary results, these models sacrificed fairness and were discarded on these grounds. Ensembles were not selected as they were found to amplify the effects of domain shift, especially considering the low prevalence of pCR.

There are certain limitations to our study. On the segmentation side, a decline in performance is observed for younger post-menopausal women (ages 0–40). This shortfall can probably be attributed to the under-representation of that subgroup in our training set. Results on the held-out validation set of the challenge (which is not publicly-available) for the pCR classification model are not available due to technical reasons. Regarding classification, the model was only exposed to radiomics features calculated from the cropped area of the malignant lesions. As such, context of the areas surrounding the lesions was not available and patient characteristics were less deductible. Overall, there was a large domain shift between the training and test data, which was advantageous for evaluating robustness of methods, but can had a large impact on performance, especially on classification tasks were population composition might play a role.

Future work can benefit from increasing the size and diversity of training datasets, potentially with private data, in order to alleviate the performance gap across all stratified cohorts. Integration of fairness-related features (e.g. age, breast density, menopausal status), either by making them available to the model or inferring them using machine learning techniques, would potentially elevate the method's ability to adjust outputs based on the patient's characteristics,

[1] The test set results reported here are preliminary estimates derived from intermediate output plots.

leading to more patient-aware models and more personalized healthcare integrations. Addressing both tasks simultaneously with a common multi-modal architecture could further be explored, either through means of adapting the segmentation convolutional neural network or exploring transformer-based architectures, which could potentially expose the model to more context and abstract features.

5 Conclusion

In this work, we developed reproducible machine learning models for breast cancer segmentation and pCR prediction using only publicly-available breast MRI data. Our segmentation approach, leveraging large-scale self-supervised pretraining, achieved top performance in the MAMA-MIA challenge validation phase (results ranking pending at the time of writing), with strong generalization and fairness across domains. The classification model maintained reasonable fairness and highlighted the potential of radiomics-based pCR prediction, especially when adequate quality segmentations can be used to delineate the areas of interest. These results demonstrate the viability of building robust and fair breast MRI models without private data, paving the way for more accessible and equitable artificial intelligence in medical imaging. Source code, data splits, and pre-training examination IDs will be made available before the workshop.

References

1. Bray, F., et al.: Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries.CA: Can. J. Clin. **74**(3), 229–263 (2024)
2. Lo Gullo, R., et al.: AI applications to breast MRI: Today and tomorrow. J. Magn. Reson. Imaging **60**(6), 2290–2308 (2024)
3. Garrucho, L., et al.: A large-scale multicenter breast cancer DCE-MRI benchmark dataset with expert segmentations. Sci. Data **12**(1), 453 (2025). https://doi.org/10.1038/s41597-025-04707-4
4. Stefanicka-Wojtas, D., Kurpas, D.: Barriers and facilitators to the implementation of personalised medicine across Europe. J. Personalized Med. **13**(2), 203 (2023)
5. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat. Methods **18**(2), 203–211 (2021)
6. Kuhl, C.K.: Abbreviated breast MRI: state of the art. Radiology **310**(3), e221822 (2024)
7. Isensee, F., et al.: NnU-net revisited: a call for rigorous validation in 3D medical image segmentation. In: Lecture Notes in Computer Science. LNCS, pp. 488–498 (2024)
8. He, K., Chen, X., Xie, S., Li, Y., Dollar, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 16000–16009 (2022)
9. Wald, T., et al.: Revisiting MAE pre-training for 3D medical image segmentation (Version 3). arXiv. https://doi.org/10.48550/ARXIV.2410.23132 (2024)

10. Ulrich, C., Isensee, F., Wald, T., Zenk, M., Baumgartner, M., Maier-Hein, K.H.: MultiTalent: a multi-dataset approach to medical image segmentation. In: Greenspan, H., et al. (ed.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2023 (2023)
11. Griethuysen, J.J.M., et al.: Computational radiomics system to decode the radiographic phenotype. Can. Res. **77**(21), e104–e107 (2017)
12. Pedregosa, F., et al.: Scikit-learn: machine learning in python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
13. Houssami, N., et al.: Meta-analysis of the association of breast cancer subtype and pathologic complete response to neoadjuvant chemotherapy. Eur. J. Can. **48**(18), 3342–3354 (2012)
14. Clark, K., et al.: The cancer imaging archive (TCIA): maintaining and operating a public information repository. J. Digit. Imaging **26**(6), 1045–57 (2013). https://doi.org/10.1007/s10278-013-9622-7. PMID: 23884657; PMCID: PMC3824915
15. Daniels, D., Last, D., Cohen, K., Mardor, Y.,Sklair-Levy, M.: Standard and delayed contrast-enhanced MRI of malignant and benign breast lesions with histological and clinical supporting data (advanced-MRI-breast-lesions) [Data set]. https://doi.org/10.7937/C7X1-YN57 (2024)
16. Saha, A., et al.: Dynamic contrast-enhanced magnetic resonance images of breast cancer patients with tumor locations [Data set]. https://doi.org/10.7937/TCIA.E3SV-RE93 (2021)
17. Newitt, D., Hylton, N.: Multi-center breast DCE-MRI data and segmentations from patients in the I-SPY 1/ACRIN 6657 trials [Data set]. https://doi.org/10.7937/K9/TCIA.2016.HDHPGJLK (2016)
18. Comstock, C. E., et al.: Abbreviated breast MRI and digital tomosynthesis mammography in screening women with dense breasts (EA1141) [Data set]. https://doi.org/10.7937/2BAS-HR33 (2023)
19. Kinahan, P., Muzi, M., Bialecki, B., Herman, B., Coombs, L.: ACRIN-contralateral-breast-MR (ACRIN 6667) [Data set]. https://doi.org/10.7937/Q1EE-J082 (2021)



Input Simplification Impact on Robustness for Targeted Therapy Subtypes in Breast MRI Segmentation AI

Alba Bernal Rodriguez[1(✉)], Beatriz Remeserio[2], Justin Engelmann[3], Lucas Gago[4], and Jacinto Velasco Rebolledo[1]

[1] Universidad Alfonso X el Sabio, Villanueva de la Cañada, Spain
albabr08@gmail.com,jvelareb80@gmail.com
[2] Universidad de Oviedo, Oviedo, Spain
[3] University of Barcelona, Barcelona, Spain
[4] Universidad de Barcelona, Barcelona, Spain

Abstract. Multi-channel approaches in medical imaging AI prioritize accuracy, but their impact on algorithmic robustness across clinically critical subgroups remains underexplored. We present a systematic comparison between single-channel (post-contrast) and multi-channel (pre+post+subtraction) strategies using the 306 test cases from the MAMA-MIA Challenge dataset, which spans four institutions. Despite similar overall performance (Dice: 0.743 vs 0.739, $p = 0.67$), single-channel models exhibited superior robustness for molecular subtypes that require targeted therapies. HR-/HER2+ cases improved across all five robustness metrics (Cohen's $d = 0.320$), while HR+/HER2+ improved in 4/5 metrics. For patients receiving anti-HER2 therapies costing over \$100,000/year ($n = 93$), single-channel segmentation achieved reliable detection (Dice ≥ 0.5) in 82.8% of cases versus 80.6% with multi-channel. Bootstrap validation ($n = 10{,}000$) confirmed these advantages, with 95% confidence intervals supporting consistency improvements. These findings challenge the assumption that greater input complexity improves clinical reliability, demonstrating that input simplification can enhance robustness and guide equitable AI deployment in precision oncology.

Keywords: Robustness in AI · Breast MRI · Input simplification · Precision oncology · Algorithmic consistency

1 Introduction

Multi-channel approaches in medical imaging AI have become the default paradigm, based on the assumption that incorporating more input sequences necessarily leads to superior and more reliable performance [1]. This has fueled the development of increasingly complex architectures integrating multiple imaging phases, contrast sequences, and derived representations [2]. However, while optimization of mean accuracy has dominated research, algorithmic robustness across clinically meaningful subgroups remains critically underexplored [3].

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
T. Zhang et al. (Eds.): Deep-Breath 2025, LNCS 16142, pp. 320–328, 2026.
https://doi.org/10.1007/978-3-032-05559-0_32

In breast cancer precision oncology, molecular subtype classification informs treatment decisions with profound clinical and economic consequences [4]. For example, HR+/HER2+ patients receive dual anti-HER2 therapy (trastuzumab plus pertuzumab) at an annual cost of approximately $100,000–150,000 [5], while HR-/HER2+ patients–typically representing HER2-enriched tumors–undergo similarly aggressive and expensive regimens [6]. In this context, consistency and reliability of algorithmic outputs are as critical as peak performance, yet current evaluation frameworks focus almost exclusively on mean performance metrics [7].

Dynamic contrast-enhanced MRI (DCE-MRI) plays a central role in the diagnosis, staging, and treatment monitoring of breast cancer, particularly for assessing response to targeted therapies [8]. Automated and accurate segmentation of primary tumors facilitates quantitative evaluation, guiding therapeutic decisions. Despite this clinical relevance, the relationship between input complexity and robustness of segmentation performance for distinct molecular subtypes remains poorly characterized.

While fairness research in AI has increasingly addressed demographic and institutional biases [9], the impact of model complexity on robustness across clinically actionable subtypes has not been systematically studied. This gap is particularly concerning in precision oncology, where inconsistency for certain subgroups may translate into suboptimal treatment planning or inequitable access to care.

To address this, we present the first systematic analysis of input complexity versus algorithmic robustness in breast MRI segmentation, focusing on molecular subtypes with direct therapeutic implications. Our study challenges the prevailing notion that greater input complexity always yields better clinical utility and provides actionable guidance for designing robust, equitable AI tools in oncology.

2 Methods

2.1 Dataset and Study Population

We utilized the official test split from the MAMA-MIA challenge dataset [10], comprising 306 patients across four medical institutions: DUKE (91 cases, 29.7%), ISPY2 (131 cases, 42.8%), ISPY1 (67 cases, 21.9%), and NACT (17 cases, 5.6%). The dataset includes comprehensive clinical metadata, including demographic characteristics (age, ethnicity), molecular subtypes (HR, ER, PR, HER2 status), treatment responses (pathological complete response - pCR), and technical parameters (scanner manufacturer, field strength).

All cases included pre-treatment T1-weighted dynamic contrast-enhanced MRI with expert tumor annotations. Molecular subtype classifications were based on immunohistochemistry and in-situ hybridization results following standard clinical guidelines.

2.2 Model Architecture and Input Strategies

We define two segmentation approaches using nnU-Net v2 [11], both trained with identical procedures to ensure a fair comparison:

- **Single-channel input:** A single post-contrast DCE-MRI volume exclusively.
- **Multi-channel input:** A three-channel strategy composed of:
  - Channel 0: Pre-contrast DCE-MRI volume
  - Channel 1: Post-contrast DCE-MRI volume
  - Channel 2: Subtraction image (Post-contrast minus Pre-contrast)

Each input channel consisted of a single 3D volume rather than a full temporal series. Specifically, we selected the first post-injection dynamic phase as the representative post-contrast image (typically acquired around 90 s after gadolinium administration), and the baseline scan as the pre-contrast image. The subtraction channel was generated voxel-wise by subtracting the pre-contrast from the post-contrast volume. This configuration mirrors common clinical protocols and ensures consistent spatial alignment across channels while avoiding the complexity and variability introduced by full multi-phase modeling.

Both models were trained using the full-resolution 3D configuration of nnU-Net and its automated preprocessing pipeline, including channel-wise z-score normalization. Although nnU-Net is configured for five-fold cross-validation by default, training and validation were performed using only fold 0 of the official training split. This choice was made to reduce computational demands during initial benchmarking; future work will expand to full cross-validation for enhanced robustness.

2.3 Robustness Evaluation Framework

We developed a comprehensive robustness assessment incorporating five complementary metrics to evaluate algorithmic consistency:

1. **Coefficient of Variation (CV):** Standard deviation divided by mean, providing scale-invariant consistency measurement
2. **Interquartile Range (IQR):** 75th–25th percentile difference, robust to outliers
3. **Median Absolute Deviation (MAD):** Median-based variability measure
4. **90th–10th Percentile Range:** Robust range excluding extreme values
5. **Clinical Threshold Analysis:** Proportion of cases below Dice = 0.5 (poor clinical utility)

For each molecular subtype, we computed robustness improvements as the difference between multi-channel and single-channel metrics (positive values indicate single-channel superiority).

2.4 Statistical Analysis and Bootstrap Validation

Statistical significance was assessed through paired t-tests for normally distributed metrics and Wilcoxon signed-rank tests for non-parametric comparisons. Effect sizes were quantified using Cohen's d for clinical interpretation.

Robustness metrics were validated through bootstrap resampling (n = 10,000 iterations) with 95% confidence intervals. For each bootstrap iteration, cases were resampled with replacement, and robustness metrics were recalculated to assess the stability of our findings.

Clinical impact was quantified through threshold analysis, examining the proportion of cases achieving clinically reliable segmentation (Dice ≥ 0.5) for molecular subtypes requiring targeted therapies.

3 Results

3.1 Overall Model Performance

Across all 306 test cases, the comparative analysis revealed statistically equivalent performance between single-channel (A2) and multi-channel (A3) approaches (Fig. 1a). The single-channel model achieved a mean Dice score of 0.743 ± 0.220 (median: 0.813, range: 0.000–0.973), while the multi-channel model achieved 0.739 ± 0.215 (median: 0.813, range: 0.000–0.973). Paired t-test analysis confirmed no significant difference ($p = 0.670$, Cohen's $d = 0.024$), establishing performance equivalence as the foundation for robustness analysis.

3.2 Input Strategy Comparison Analysis

Table 1 shows the detailed performance comparison between approaches across molecular subtypes. While overall performance was equivalent, distinct patterns emerged for specific molecular subtypes requiring targeted therapies, as visualized in Fig. 1b.

Table 1. Performance by Input Strategy and Molecular Subtype

Molecular Subtype	N	Dice Mean (Single)	CV (Single)	Dice Mean (Multi)	CV (Multi)	Robustness Advantage
HR+/HER2+	58	0.696 ± 0.254	0.364	0.669 ± 0.268	0.401	Single-channel
HR+/HER2-	115	0.741 ± 0.211	0.284	0.752 ± 0.183	0.243	Multi-channel
HR-/HER2+	35	0.779 ± 0.165	0.212	0.747 ± 0.197	0.264	Single-channel
Triple-negative	92	0.772 ± 0.202	0.262	0.768 ± 0.206	0.269	Single-channel

CV = Coefficient of Variation (lower values indicate better consistency);
N = number of subtype samples;
Robustness Advantage indicates which approach showed superior consistency metrics.

Table 2. Robustness Analysis by Molecular Subtype Single-channel Advantages

Molecular Subtype	N	CV Imp.	IQR Imp.	MAD Imp.	Score
HR+/HER2+	58	+0.036	+0.028	–0.011	4/5
HR+/HER2	115	–0.041	0.037	–0.014	0/5
HR/HER2+	35	+0.052	+0.089	+0.029	5/5
Triple-negative	92	+0.007	–0.041	–0.003	1/5

CV Imp. = Coefficient of Variation improvement; IQR Imp. = Interquartile Range improvement; MAD Imp. = Median Absolute Deviation improvement; Score = Number of metrics favoring single-channel approach out of 5 total metrics.

3.3 Robustness Analysis Across Molecular Subtypes

Table 2 presents a comprehensive robustness analysis across five complementary metrics. HR-/HER2+ cases demonstrated superior robustness with single-channel approaches across all evaluated metrics, achieving a perfect 5/5 robustness score with medium effect size (Cohen's d = 0.320, p = 0.067), as detailed in Fig. 2.

HR+/HER2+ cases (n = 58) showed robustness improvements in 4/5 metrics. Notably, both HR+/HER2+ and HR-/HER2+ subtypes represent patients requiring targeted anti-HER2 therapies costing over $100,000 annually, highlighting the clinical significance of these consistency improvements.

These robustness differences may reflect the underlying enhancement patterns characteristic of each molecular subtype. For instance, HR-/HER2+ tumors–often classified as HER2-enriched–typically exhibit rapid and heterogeneous contrast uptake, with irregular margins and high spatial variability. This complexity may introduce noise when combining multiple sequences, potentially degrading segmentation consistency. In contrast, single-phase post-contrast inputs may offer a cleaner and more focused representation of the tumor, capturing peak enhancement without incorporating redundant or conflicting signals.

On the other hand, HR+/HER2 tumors often present with slow, homogeneous enhancement patterns and well-defined borders, making them more amenable to multichannel fusion strategies. These observations suggest that the effectiveness of input simplification may depend on the biological and radiological properties of each subtype, supporting a more nuanced approach to model design in precision oncology.

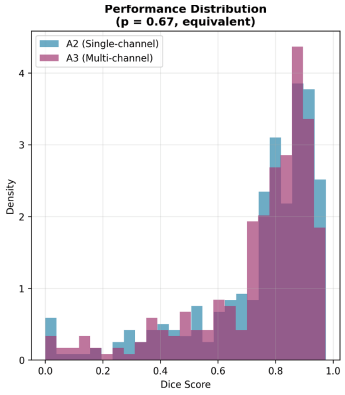
3.4 Clinical Impact and Treatment Decision Analysis

For patients requiring targeted anti-HER2 therapies (n=93), single-channel approaches provided reliable tumor detection (Dice $\geq$ 0.5) in 82.8% of cases compared to 80.6% with multi-channel segmentations. This represents 2 additional patients with clinically actionable segmentation quality, translating to improved treatment planning reliability for high-cost therapeutic interventions.

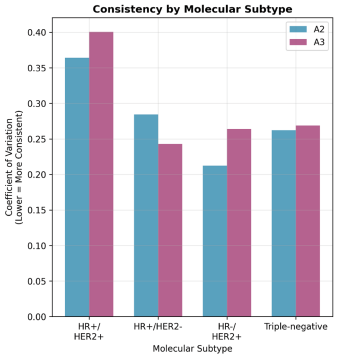
Table 3. Bootstrap Validation Summary - Robustness Improvements (n = 10,000)

Molecular Subtype	CV Improvement	95% CI	p-value	Clinical Impact
HR+/HER2+	+0.0367	[-0.066, 0.143]	0.249	Targeted therapy required
HR+/HER2-	-0.0416	[-0.093, 0.004]	0.961	Standard luminal treatment
HR-/HER2+	+0.0519	[-0.012, 0.115]	0.061	HER2-enriched, aggressive
Triple-negative	+0.0064	[-0.036, 0.051]	0.392	Chemotherapy protocol

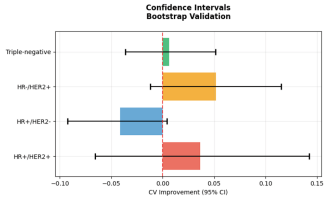
CV Improvement = Mean coefficient of variation improvement across bootstrap iterations; CI = Confidence Interval; p-value = proportion of iterations where single-channel was not superior.



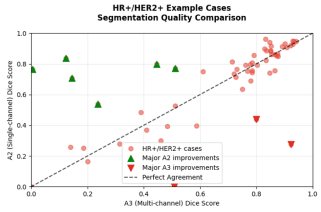
(a) Performance distribution equivalence (p = 0.67)



(b) Consistency by molecular subtype

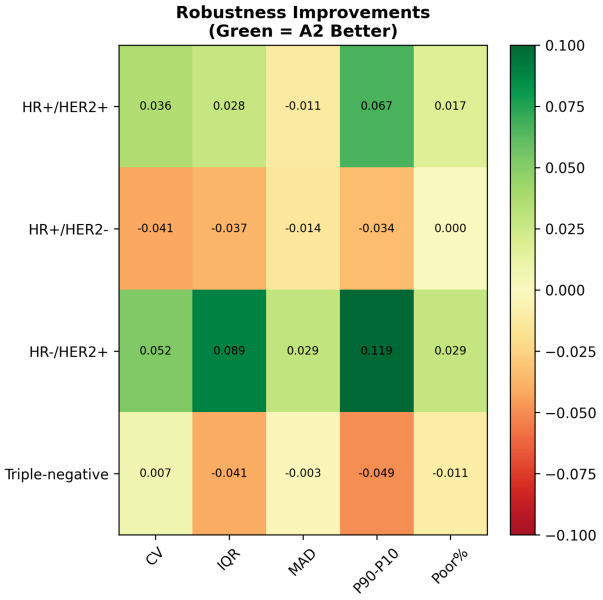


(c) Bootstrap confidence intervals

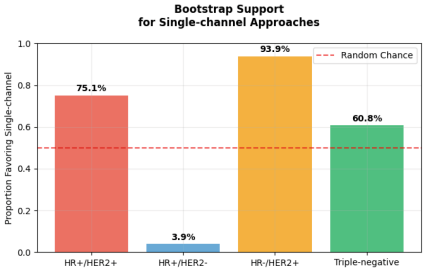


(d) HR+/HER2+ case-by-case comparison

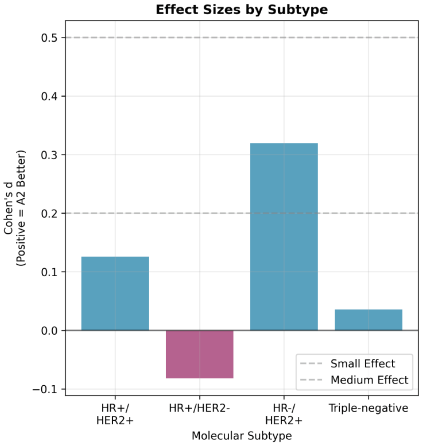
Fig. 1. Main findings of input strategy comparison. (a) Overall performance distributions showing statistical equivalence between single-channel and multi-channel approaches. (b) Coefficient of variation by molecular subtype, demonstrating superior consistency for HR+/HER2+ and HR-/HER2+ with single-channel approaches. (c) Bootstrap validation confidence intervals confirming robustness improvements. (d) Individual case analysis for HR+/HER2+ subtype showing distribution of improvements (green triangles) and degradations (red triangles). (Color figure online)



(a) Comprehensive robustness improvements



(b) Bootstrap support analysis



(c) Effect sizes by subtype

Fig. 2. Detailed robustness analysis and validation. (a) Heatmap showing improvements across five robustness metrics for each molecular subtype (green indicates single-channel superiority, with HR-/HER2+ achieving a perfect 5/5 score). (b) Bootstrap support showing proportion of 10,000 iterations favoring single-channel approaches, with HR-/HER2+ achieving 93.9% support and HR+/HER2+ achieving 75.1%. (c) Effect sizes (Cohen's d) demonstrating a medium effect for HR-/HER2+ subtype (0.320) and a small effect for HR+/HER2+ (0.126). (Color figure online)

Among HR+/HER2+ patients specifically, single-channel approaches showed net clinical benefit with 10 cases demonstrating clinically significant improvement (Dice difference >0.1) versus 8 cases with degradation. Cases with very poor performance (Dice<0.3) were reduced from 8 to 7 for the HR+/HER2+ subtype and from 1 to 0 for the HR-/HER2+ subtype.

3.5 Bootstrap Validation and Statistical Robustness

Bootstrap validation (n = 10,000 iterations) confirmed the consistency of robustness improvements (Table 3). For HR-/HER2+ cases, 93.9% of bootstrap iterations demonstrated single-channel consistency advantages (95% CI: [–0.012, 0.115]). HR+/HER2+ cases showed consistent improvement trends with 75.1% of iterations favoring single-channel approaches (95% CI: [–0.066, 0.143]).

4 Conclusion

Our findings challenge the common assumption that increased input complexity improves clinical AI performance. Despite equivalent overall Dice scores, single-channel models consistently demonstrated superior robustness across molecular subtypes requiring targeted therapies–particularly HR+/HER2+ and HR-/HER2+–with improvements confirmed by five robustness metrics and bootstrap validation. These results suggest that additional input channels may introduce correlated noise, especially in subtypes with heterogeneous enhancement patterns.

Clinically, these findings offer immediate guidance: single-channel approaches enhance segmentation consistency without performance loss, supporting more reliable treatment decisions in high-cost scenarios such as anti-HER2 therapy. The observed 2.2% improvement in reliable detection (82.8% vs 80.6%) carries significant value in precision oncology, where segmentation quality can directly impact treatment monitoring and planning.

From an implementation standpoint, single-channel models reduce computational cost, simplify preprocessing, and facilitate deployment in low-resource settings. These operational benefits, combined with superior robustness, make them highly attractive for real-world clinical use.

Furthermore, our results extend algorithmic fairness beyond demographic parity to include consistency across clinically meaningful subgroups. The fact that simplification disproportionately benefits the most challenging subtypes highlights its potential as a bias mitigation strategy, reinforcing equity in treatment outcomes.

A key limitation of this work is the small sample size for HR-/HER2+ cases (n = 35), which may limit statistical power. Nevertheless, the consistency of results across all metrics and subtypes lends confidence to our conclusions. Expanding datasets–especially for underrepresented subgroups–will be essential to validate and extend these findings.

Overall, input simplification appears to serve dual purposes in breast MRI segmentation AI: enhancing both performance stability and fairness. We advocate for its broader consideration in clinical AI development, particularly in applications prioritizing consistency and equity in high-stakes decision-making.

To complement the quantitative results presented here, future work will include qualitative case analyses, particularly of those that showed significant improvement with input simplification or when comparing multichannel variants.

References

1. Litjens, G., et al.: A survey on deep learning in medical image analysis. Med. Image Anal. **42**, 60–88 (2017)
2. Zhou, T., et al.: A review: deep learning for medical image segmentation using multi-modality fusion. Array **3**, 100004 (2019)
3. Chen, I.Y., Pierson, E., Rose, S., Joshi, S., Ferryman, K., Ghassemi, M.: Ethical machine learning in healthcare. Annu. Rev. Biomed. Data Sci. **4**, 123–144 (2021)
4. Slamon, D., et al.: Adjuvant trastuzumab in HER2-positive breast cancer. N. Engl. J. Med. **365**(14), 1273–1283 (2011)
5. Institute for Clinical and Economic Review: Anti-HER2 therapy costs and effectiveness. ICER Evidence Report (2023)
6. Loibl, S., et al.: HER2-positive breast cancer. Lancet **389**(10087), 2415–2429 (2021)
7. Rajkomar, A., Hardt, M., Howell, M.D., Corrado, G., Chin, M.H.: Ensuring fairness in machine learning to advance health equity. Ann. Intern. Med. **169**(12), 866–872 (2018)
8. Mann, R.M., Cho, N., Moy, L.: Breast MRI: state of the art. Radiology **292**(3), 520–536 (2019)
9. Seyyed-Kalantari, L., Zhang, H., McDermott, M.B., Chen, I.Y., Ghassemi, M.: Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nat. Med. **27**(12), 2176–2182 (2021)
10. Garrucho, L., et al.: A large-scale multicenter breast cancer DCE-MRI benchmark dataset with expert segmentations. Sci. Data **12**, 453 (2025)
11. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat. Methods **18**(2), 203–211 (2021)
12. Sutton, E.J., et al.: Breast cancer molecular subtype classifier that incorporates MRI features. J. Magn. Reson. Imaging **44**(1), 122–129 (2020)
13. Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. Adv. Neural. Inf. Process. Syst. **29**, 3315–3323 (2016)

LesiOnTime - Joint Temporal and Clinical Modeling for Small Breast Lesion Segmentation in Longitudinal DCE-MRI

Mohammed Kamran[1,2], Maria Bernathova[3], Raoul Varga[4], Christian F. Singer[5,7], Zsuzsanna Bago-Horvath[6,7], Thomas Helbich[4], Georg Langs[1,2], and Philipp Seeböck[1,2(✉)]

[1] Computational Imaging Research Lab, Department of Biomedical Imaging and Image-guided Therapy, Medical University of Vienna, Vienna, Austria
{mohammed.kamran,philipp.seeboeck}@meduniwien.ac.at
[2] Comprehensive Center for AI in Medicine, Medical University of Vienna, Vienna, Austria
[3] Breast Health Center, Privatklinik Döbling, Vienna, Austria
[4] Department of Biomedical Imaging and Image-guided Therapy, Medical University of Vienna, Vienna, Austria
[5] Department of Obstetrics and Gynecology, Medical University of Vienna, Vienna, Austria
[6] Department of Pathology, Medical University of Vienna, Vienna, Austria
[7] Comprehensive Cancer Center, Medical University of Vienna, Vienna, Austria

Abstract. Accurate segmentation of small lesions in Breast Dynamic Contrast-Enhanced MRI (DCE-MRI) is critical for early cancer detection, especially in high-risk patients. While recent deep learning methods have advanced lesion segmentation, they primarily target large lesions and neglect valuable longitudinal and clinical information routinely used by radiologists. In real-world screening, detecting subtle or emerging lesions requires radiologists to compare across timepoints and consider previous radiology assessments, such as the BI-RADS score. We propose *LesiOnTime*, a novel 3D segmentation approach that mimics clinical diagnostic workflows by jointly leveraging longitudinal imaging and BI-RADS scores. The key components are: (1) a *Temporal Prior Attention (TPA)* block that dynamically integrates information from previous and current scans; and (2) a *BI-RADS Consistency Regularization (BCR)* loss that enforces latent space alignment for scans with similar radiological assessments, thus embedding domain knowledge into the training process. Evaluated on a curated in-house longitudinal dataset of high-risk patients with DCE-MRI, our approach outperforms state-of-the-art single-timepoint and longitudinal baselines by 5% in terms of Dice. Ablation studies demonstrate that both TPA and BCR contribute complementary performance gains. These results highlight the importance of incorporating temporal and clinical context for reliable early lesion segmentation in real-world breast cancer screening. Our code is publicly available at https://github.com/cirmuw/LesiOnTime.

© The Author(s) 2026
T. Zhang et al. (Eds.): Deep-Breath 2025, LNCS 16142, pp. 329–339, 2026.
https://doi.org/10.1007/978-3-032-05559-0_33

Keywords: Small Lesion Segmentation · Breast DCE-MRI · Longitudinal Modeling · Clinical Priors

1 Introduction

Breast cancer is the most prevalent cancer among females, with about 2.3 million diagnosed cases and 670,000 deaths reported globally in 2022 alone [19]. In the United States alone, an estimated number of 316,950 new invasive breast cancer cases are projected for 2025 [1]. Early detection significantly improves the 5-year survival rate, up to 90% [18].

2D Mammography-based screening is widely accepted as an effective technique for early detection [10]. However, for individuals at elevated risk–such as those with a family history or pathogenic mutations conferring a lifetime risk of >30%, 3D Dynamic Contrast-Enhanced MRI (DCE-MRI) is recommended due to its superior sensitivity [13]. Radiologists assess DCE-MRI using the Breast Imaging-Reporting and Data System (BI-RADS) (0–6) to estimate malignancy likelihood and guide follow-up, where scores ≤ 3 suggest benignity, and ≥ 4 indicate malignancy probabilities between 2% and 95%, warranting further workup. However, accurate detection of small or subtle lesions (median volume of 800 mm^3 or even less) in DCE-MRI remains challenging due to varying size, shape, enhancement patterns and low contrast with adjacent tissues. Normal background parenchymal enhancement (BPE) can mimic or mask lesions, especially in premenopausal women. There is also significant intra- and inter-reader variability, even among experienced radiologists with years of training. In addition, inconsistent acquisition protocols and a lack of large, annotated datasets hinder the development of robust AI segmentation systems.

Related Work. The availability of annotated breast DCE-MRI datasets has driven progress in automated lesion analysis [4,5,17,22]. While early work focused on 2D slice-wise tumor segmentation [3,20], recent approaches leverage 3D models [5,22], contrastive learning [6], and diffusion models [12]. GAN-based anomaly detection has shown promise for identifying subtle early changes in high-risk women, even before visible lesions emerged [2]. Along the same line of research, pre-diagnostic scans from high-risk cohorts up to one year before cancer identification have shown predictive potential [8], highlighting the importance of longitudinal data. State-of-the-art segmentation models include CNN-based nnUNet [9] and transformer-based architectures such as SwinUNETR [7] and SAM [11]. Longitudinal extensions like *LongiSeg* [16] have been proposed in other domains (e.g., multiple sclerosis lesion segmentation in brain MRI), but adaptation to breast MRI remains largely unexplored. Moreover, most studies still target large lesions in non-high-risk populations [5,21,22] and rely on single time-point segmentation models [7,9,11], neglecting longitudinal context from prior scans [4,5,22] or clinical information such as BI-RADS scores [16]–data routinely available in clinical follow-ups. While *LongiSeg* [16] represents a step toward longitudinal segmentation, it requires the availability of manual

pixel-wise annotations of previous timepoints, which are typically unavailable in breast screening scenarios where lesions are absent or undetectable at earlier timepoints. Moreover, it ignores complementary clinical context such as BI-RADS scores, limiting its utility in real-world diagnostic workflows.

Contribution. To address these limitations, we propose *LesiOnTime*, a novel longitudinal segmentation approach for breast DCE-MRI. It integrates temporal imaging (DCE-MRI) and clinical information (BI-RADS scores) via two main innovations, the Temporal Prior Attention (TPA) Block and BI-RADS Consistency Regularizer (BCR). In contrast to [16], our approach does not require a manual pixel-wise lesion annotation for the previous timepoint, as the TPA block dynamically learns to weight the relevance of longitudinal scans. For instance, if a previous scan shows very little change compared to the current one, the network can down-weight it. Conversely, in cases where the temporal progression is meaningful (e.g., subtle enhancement appearing over time), the model learns to emphasize it. Furthermore, we incorporate clinical information during training using the BCR loss. It encourages feature similarity across timepoints only when BI-RADS scores are consistent, and allows divergence when scores indicate clinical progression–thus enforcing both temporal and clinical consistency.

Experiments on a curated high-risk cohort show that LesiOnTime surpasses state-of-the-art methods by 5% Dice, including both single-timepoint and longitudinal models (nnUNet [9], SwinUNETR [7], LongiSeg [16]). Ablation studies confirm the complementary value of TPA and BCR: removing either component reduces performance, while their combination yields the best results.

2 Method

We propose **LesiOnTime**, a novel 3D segmentation approach that jointly models temporal imaging features and clinical priors (Fig. 1). It integrates temporal information through a temporal prior attention (TPA) block (Sect. 2.1) and clinical context via a BI-RADS consistency regularization loss (BCR) (Sect. 2.2). Formally, let $x_t, x_{t-1} \in \mathbb{R}^{C\times D\times H\times W}$ denote 3D breast DCE-MRI volumes for visits at the timepoints t (current) and $t-1$ (previous), with y_t as the corresponding ground truth segmentation mask. The model f_θ, parameterized by θ, takes the pair of volumes as input and predicts the segmentation mask of the current timepoint t, $y'_t = f_\theta(x_t, x_{t-1})$.

LesiOnTime adopts a shared-weight dual encoder-decoder architecture tailored for longitudinal lesion analysis. The model processes pair of scans x_t, x_{t-1} (current and prior) through two parallel encoders with shared weights, ensuring consistent feature representations across visits. At each layer m, the encoded feature maps from the current scan k_t^m and the prior scan k_{t-1}^m are passed into a **Temporal Prior Attention (TPA)** block, which captures and modulates temporal dependencies.

The model is trained end-to-end using a composite loss: (1) the **segmentation loss** combining **Dice loss** and **Cross-Entropy loss** for accurate lesion

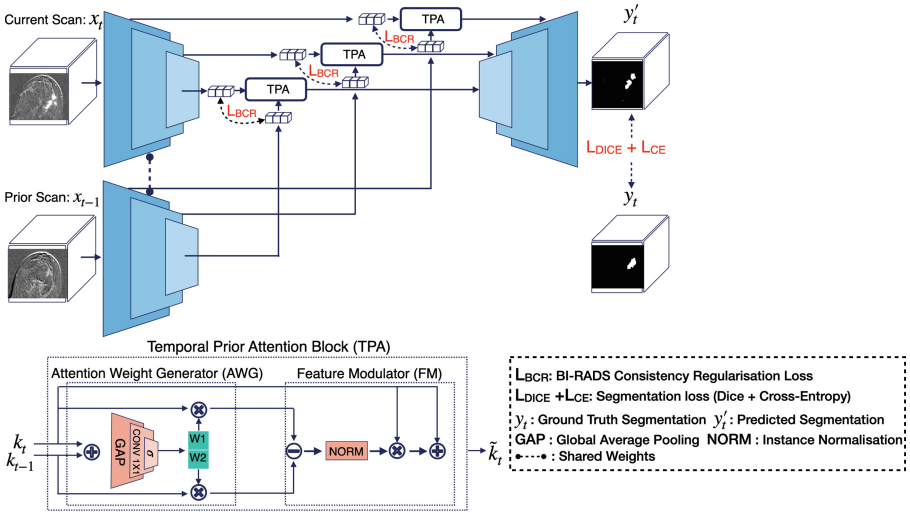


Fig. 1. LesiOnTime: Paired current and prior scans (x_t,x_{t-1}) are processed by a dual-encoder with shared weights. Features from both encoders (k_t,k_{t-1}) pass through the proposed TPA block at each skip connection, which includes an attention weight generator (AWG) and a feature modulator (FM). AWG produces attention weights (W_1, W_2) for the scan pair, used by the FM to adaptively combine features. The BCR loss ($\mathcal{L}_{\text{BCR}}$) regularizes the latent space according to BI-RADS progression during training. The predicted segmentation map (y'_t) is compared with the ground truth (y_t) using a combination of Dice and Cross-Entropy loss ($\mathcal{L}_{\text{DICE}} + \mathcal{L}_{\text{CE}}$).

delineation, and (2) the proposed **BI-RADS Consistency Regularization (BCR) loss** (Sect. 2.2), which constrains feature embeddings in accordance with clinical progression.

$$\mathcal{L}_{\text{total}} = \lambda_{dice}\mathcal{L}_{\text{Dice}} + \lambda_{ce}\mathcal{L}_{\text{CE}} + \lambda_{bcr}\mathcal{L}_{\text{BCR}} \tag{1}$$

2.1 Temporal Prior Attention (TPA) Block

The TPA block consists of two main components: the *attention weight generator* and the *feature modulator*. Given feature maps $k_t^m, k_{t-1}^m \in \mathbb{R}^{c\times d\times h\times w}$ from the dual encoders at layer m, they are concatenated along a new dimension into a new tensor of shape $\mathbb{R}^{2\times c\times d\times h\times w}$.

The AWG applies global average pooling across the spatial dimensions (D, H, W), reducing each channel to a scalar. The resulting vector is passed through a 1D convolution layer and a sigmoid activation to generate soft attention weights $w = [w_1, w_2] \in \mathbb{R}^{2\times 1}$, where w_1 and w_2 are scalar values representing the attention assigned to visits t and $t-1$, respectively.

The FM uses these learned attention weights to adaptively adjust the current feature map k_t^m in relation to the prior feature map k_{t-1}^m. This is done through

a weighted residual difference:

$$\tilde{k}_t^m = k_t^m \odot \mathrm{InstNorm}(w_1 \odot k_t^m - w_2 \odot k_{t-1}^m) + k_t^m, \tag{2}$$

where $\odot$ denotes element-wise (broadcasted) multiplication across channels. This formulation allows the model to emphasize meaningful temporal changes while preserving the core features of the current scan.

By dynamically adjusting attention across each patient's scan pair, the TPA block learns to model cross-scan interactions. This enables the network to focus on the most informative timepoint, enhancing sensitivity to subtle or slowly evolving lesions.

2.2 BI-RADS Consistency Regularizer

The BI-RADS Consistency Regularization (BCR) loss is designed to enforce **clinical and temporal coherence** in feature learning across longitudinal breast MRI scans. Specifically, it constrains the distance between feature embeddings of consecutive scans based on the change in BI-RADS scores. At each layer m, the feature-level BCR loss is defined as:

$$\mathcal{L}_{\text{feat}}^{(m)} = \frac{\tanh\left(\left\|k_t^{(m)} - k_{t-1}^{(m)}\right\|^2\right)}{\Delta\text{BI-RADS} + \epsilon}, \tag{3}$$

where ΔBI-RADS is the absolute difference in BI-RADS scores between visits t and $t-1$, and ϵ is a small positive constant to avoid division by zero. The *tanh* activation bounds feature differences in the numerator to [0,1), reducing sensitivity to extreme feature differences and promoting stable training. The overall BCR loss across a set of all layers $\mathcal{M}$ is computed as a weighted sum:

$$\mathcal{L}_{\text{BCR}} = \sum_{m \in \mathcal{M}} w_m \cdot \mathcal{L}_{\text{feat}}^{(m)}, \tag{4}$$

with w_m denoting layer-wise weights. Clinically, a stable BI-RADS score implies minimal pathological change, motivating feature similarity across timepoints. Conversely, an increase in BI-RADS typically reflects the emergence of new enhancement or growth of lesions in the current scan relative to prior imaging, promoting greater feature divergence. The BCR loss explicitly encodes this behavior, as it approaches zero when the BI-RADS scores are identical and the embeddings are well aligned, enabling joint modeling of longitudinal imaging and radiological assessment.

3 Experimental Setup

Material. We utilize a curated in-house breast DCE-MRI dataset comprising 155 studies (BI-RADS 4/5) acquired at the Vienna General Hospital (AKH), each with pre-contrast and at least three post-contrast T1-weighted volumes

with spatial resolutions of $48 \times 384 \times 384$ or $80 \times 512 \times 512$ and a median lesion volume of 800 mm^3. Manual pixel-wise lesion annotations were created by an expert radiologist for all 155 visits, using ITK-SNAP on the first post-contrast volume. Each study has a prior scan 6–24 months apart, enabling longitudinal modeling (310 visits total). BI-RADS scores are available for all visits, generated by radiologists in clinical routine. Data are randomly split by patient into development (n = 131, 5-fold cross-validation) and an independent test set (n = 24) for final evaluation.

Preprocessing. All volumes are resampled to a 0.7 mm × 0.7 mm × 2 mm voxel spacing. We perform rigid and nonlinear registration of the current and previous timepoint volume using the Symmetric Normalization (SyN) algorithm of Advanced Normalization Tools Python (ANTsPy). Subsequently, we crop the volumes to the breast tissue area using a breast segmentation algorithm [15], excluding thoracic structures that may confound lesion segmentation.

Experiments. We compare *LesiOnTime* to LongiSeg [16], nnUNet [9], SwinUNETR [7] (trained on 3D volumes), and BiomedParse [21], a large 2D vision-language segmentation model (fine-tuned on 2D slices with prompts). All models are evaluated consistently in 3D, with the entire volume treated as single sample. Ablation experiments include $LesiOnTime_{w/oTPA}$ (TPA replaced by difference-weighting of LongiSeg) and $LesiOnTime_{w/o\ BCR}$ (BCR loss removed).

Training Details. All models are trained on an NVIDIA RTX 6000 GPU (48 GB VRAM) using a batch size of 2 for 1000 epochs. Loss combines Dice, Cross-Entropy ($\lambda_{\text{dice}} = \lambda_{\text{ce}} = 1.0$), and BCR terms ($\lambda_{\text{fc}} = 0.1$). BCR weights w_m are linearly scaled across network layers (0.1 to 1.0) with $\epsilon = 1e{-}1$, emphasizing deeper, semantically richer features (e.g. 0.1,0.3,0.5,0.7,0.9,1 for a 6-layer network). Our longitudinal approach builds on the previously proposed architectural selection strategy to identify the most suitable U-Net configuration [9,16]. The final architecture is a 3D PlainConvUNet from the dynamic network architectures library, with 6 stages in the encoder progressively increasing feature channels [32, 64, 128, 256, 320, 320] using 3D convolutions (Conv3d) and instance normalization with a kernel size of $3 \times 3 \times 3$. The decoder mirrors this structure with 2 convolutional layers per stage. LeakyReLU is used as the nonlinearity, and the architecture avoids dropout.

4 Results and Discussion

Table 1 summarizes segmentation performance on the held-out test set. *LesiOnTime* achieves the highest Dice score (0.356), outperforming both longitudinal (LongiSeg: 0.302) and single-timepoint models (nnUNet: 0.291, SwinUNETR: 0.195, BiomedParse: 0.122). LesiOnTime also yields superior precision, recall, and HD-95 distance, indicating improved boundary delineation and lesion localization. Statistical comparison using the Wilcoxon signed-rank test revealed significant improvements in HD-95 over all baselines ($p < 0.05$), while differences

in Dice score did not reach statistical significance. We hypothesize that this may be due to the known high variability of Dice scores for small structures, where boundary-focused metrics such as HD-95 are considered more stable and sensitive to subtle segmentation improvements [14].

Table 1. Quantitative results of 3D breast lesion segmentation on the DCE-MRI test set. Best results are in bold. HD-95: 95th percentile Hausdorff distance.

Model	Dice	HD-95	Precision	Recall
nnUNET [9]	0.29	164.9	0.32	0.34
SwinUNETR [7]	0.195	182.3	0.24	0.25
BiomedParse [21]	0.12	192.2	0.14	0.20
LongiSeg [16]	0.30	132.2	0.31	0.34
LesiOnTime$_{\text{w/o TPA}}$ (ours)	0.31	110.4	0.37	0.41
LesiOnTime$_{\text{w/o BCR}}$ (ours)	0.32	112.3	0.38	0.37
LesiOnTime (ours)	**0.35**	**106.5**	**0.39**	**0.42**

This is in line with qualitative results (Fig. 2), where *LesiOnTime* shows superior delineation of lesions, particularly for extremely small lesions and non-mass enhancements with diffuse boundaries – a persistent challenge due to low contrast and annotation ambiguity. We observed that *LesiOnTime* seems to focus on areas of highest intensity and structural consistency, while radiologists may intentionally annotate larger regions to prioritize high sensitivity in ambiguous cases.

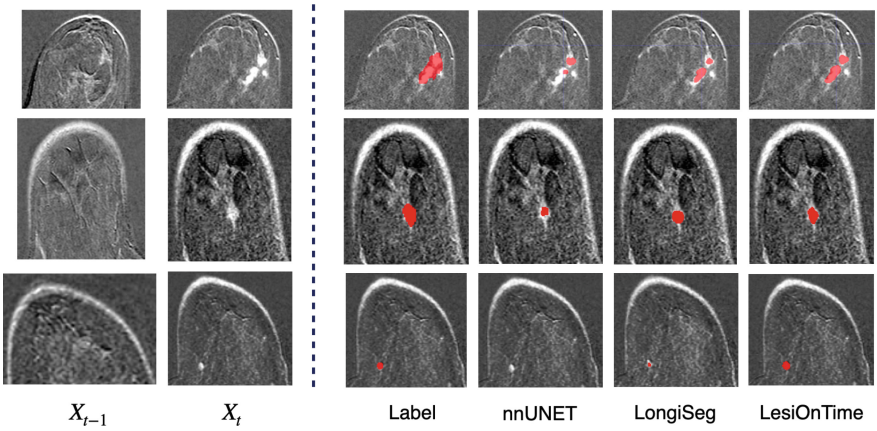


Fig. 2. Qualitative segmentation results comparing *LesiOnTime* with state-of-the-art baselines: single timepoint nnUNet [9] and multi-timepoint LongiSeg [16].

Figure 3 shows that BCR guides latent features to cluster according to BI-RADS scores. Specifically, distances between current and prior timepoints shrink for the majority of cases when trained with BCR (LesiOnTime), indicating better alignment in latent space. Without BCR (LongiSeg [16]), these distances are generally larger, underscoring BCR's role in embedding clinically relevant priors.

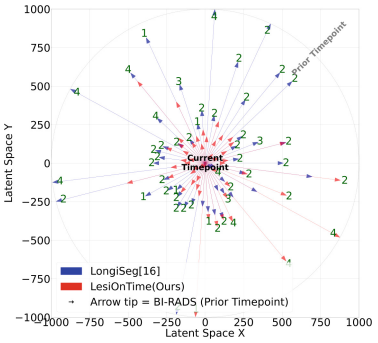


Fig. 3. This radial plot illustrates vector embeddings of current (center) and prior (outer circle) timepoints in latent space, with arrow magnitudes representing the Euclidean distance between these embeddings. It compares the effect of training with BCR (**LesiOnTime**, red) versus without BCR (LongiSeg, blue). Arrow tips are labeled with BI-RADS scores, indicating clinical suspicion at the prior scan, while all current timepoints have a fixed BI-RADS score of 4.

One limitation of this study is its single-center setting. Currently, no publicly available longitudinal breast DCE-MRI dataset exists for external validation. However, the method's utilization of temporal and radiological priors supports transferability to other sites and scanners. Notably, *LesiOnTime* requires no manual lesion annotation for prior timepoints and uses BI-RADS scores only during training, not inference, simplifying clinical deployment. While segmentation of small, diffuse lesions remains challenging (e.g. due to low uptake contrast or scattered nature), leveraging prior timepoints partially mitigates this, as shown by improved segmentation over single-timepoint baselines. This highlights the advantage of incorporating temporal context for lesion detection in screening populations. We excluded public datasets such as BreastDM [22], MAMA-MIA [5], DUKE MRI [17] and Advanced-MRI-Breast-Lesions [4] due to their focus on large or late-stage lesions and the lack of longitudinal scans, unrepresentative of our clinical scenario focused on subtle early-stage lesions. Although registration accuracy was visually validated in a subset of cases, minor misalignments cannot be fully excluded and may contribute to residual segmentation errors. However, the TPA mechanism was designed to reduce sensitivity to such discrepancies by weighting temporal features adaptively.

5 Conclusion

In this work, we present *LesiOnTime*, a novel breast DCE-MRI segmentation approach that explicitly integrates longitudinal imaging context and radiological assessments to address key limitations of conventional single-timepoint models. Through its temporal prior attention (TPA) mechanism, *LesiOnTime* leverages prior scans without requiring manual annotations for earlier timepoints. The proposed BI-RADS Consistency Regularization (BCR) aligns latent representations with diagnostic labels, embedding clinical reasoning into the model. *LesiOnTime* consistently outperforms both longitudinal and single-timepoint baselines in segmenting small, diffuse, and clinically challenging lesions typical for high-risk screening populations. These findings highlight the value of temporal and clinical cues for improving early breast cancer segmentation in MRI, and suggest that *LesiOnTime* may support future clinical workflows by enabling more reliable and context-aware lesion assessment in longitudinal breast cancer screening.

Acknowledgements. This work was supported by the European Federation for Cancer Images (EUCAIM, Grant No. 101100633), Initiative Krebsforschung - Grant 2022 (UE77104006), and Vienna Science and Technology Fund (WWTF, PREDICTOME) [10.47379/LS20065].

Disclosure of Interests. GL: meduniwien.ac.at, mit.edu, csh.ac.at, contextflow.com; PS: meduniwien.ac.at; imperial.ac.uk; univie.ac.at; TH: imperial.ac.uk; bayer.com; bd.com; novomed.at; guerbet.com; siemens.com; hologic.de; bruker.com; ZBH: MSD, Astra Zeneca, Gilead, Stemline, Abbvie.

References

1. American cancer society: breast cancer. website, American cancer society (2025), https://www.cancer.org/cancer/types/breast-cancer.html, Accessed December 2025
2. Burger, B., Bernathova, M., Seeböck, P., Singer, C.F., Helbich, T.H., Langs, G.: Deep learning for predicting future lesion emergence in high-risk breast mri screening: a feasibility study. Eur. Radiol. Exp. **7**(1), 32 (2023)
3. Carvalho, E.D., Silva, R.R.V., Mathew, M.J., Araujo, F.H.D., De Carvalho Filho, A.O.: Tumor segmentation in breast dce-mri slice using deep learning methods. In: 2021 IEEE Symposium on Computers and Communications (ISCC), pp. 1–6. IEEE (2021)
4. Daniels, D., Last, D., Cohen, K., Mardor, Y., Sklair-Levy, M.: Standard and delayed contrast-enhanced mri of malignant and benign breast lesions with histological and clinical supporting data (advanced-mri-breast-lesions)(version 2) [data set] . the cancer imaging archive (2024)
5. Garrucho, L., et al.: A large-scale multicenter breast cancer dce-mri benchmark dataset with expert segmentations. Sci. Data **12**(1), 453 (2025)
6. Guo, S., et al.: Contrastive learning-based breast tumor segmentation in dce-mri. In: International Workshop on Machine Learning in Medical Imaging, pp. 157–165. Springer (2023)

7. He, Y., Nath, V., Yang, D., Tang, Y., Myronenko, A., Xu, D.: Swinunetr-v2: stronger swin transformers with stagewise convolutions for 3d medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 416–426. Springer (2023)
8. Hirsch, L., et al.: Early detection of breast cancer in mri using ai. Acad. Radiol. **32**(3), 1218–1225 (2025)
9. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nat. Methods **18**(2), 203–211 (2021)
10. Kebede, S.R., et al.: Dual view deep learning for enhanced breast cancer screening using mammography. Sci. Rep. **14**(1), 3839 (2024)
11. Kirillov, A., et al.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4015–4026 (2023)
12. Konz, N., Chen, Y., Dong, H., Mazurowski, M.A.: Anatomically-controllable medical image generation with segmentation-guided diffusion models. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 88–98. Springer (2024)
13. Lee, C.H., et al.: Breast cancer screening with imaging: recommendations from the society of breast imaging and the acr on the use of mammography, breast mri, breast ultrasound, and other technologies for the detection of clinically occult breast cancer. J. Am. Coll. Radiol. **7**(1), 18–27 (2010)
14. Maier-Hein, L., et al.: Metrics reloaded: recommendations for image analysis validation. Nat. Methods **21**(2), 195–212 (2024)
15. Perschy, L.: Breast cancer prediction in high-risk patients using deep learning on MR imaging. Diploma thesis, Universität Wien (2023), https://doi.org/10.25365/thesis.73974
16. Rokuss, M.R., et al.: Longitudinal segmentation of ms lesions via temporal difference weighting. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 64–74. Springer (2024)
17. Saha, A., et al.: A machine learning approach to radiogenomics of breast cancer: a study of 922 subjects and 529 dce-mri features. Br. J. Cancer **119**(4), 508–516 (2018)
18. Witowski, J., et al.: Improving breast cancer diagnostics with deep learning for mri. Sci. Transl. Med. **14**(664), eabo4802 (2022)
19. World health organization: breast cancer. Fact sheet, WHO News-room, March 2024, https://www.who.int/news-room/fact-sheets/detail/breast-cancer, Accessed December 2025
20. Zhang, J., Saha, A., Zhu, Z., Mazurowski, M.A.: Breast tumor segmentation in dce-mri using fully convolutional networks with an application in radiogenomics. In: Medical Imaging 2018: Computer-Aided Diagnosis, vol. 10575, pp. 192–196. SPIE (2018)
21. Zhao, T., et al.: Biomedparse: a biomedical foundation model for image parsing of everything everywhere all at once. arXiv preprint arXiv:2405.12971 (2024)
22. Zhao, X., et al.: Breastdm: a dce-mri dataset for breast tumor image segmentation and classification. Comput. Biol. Med. **164**, 107255 (2023)

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.





MedMask: A Self-supervised Vision Foundation Model for Breast Cancer Detection Using Mammograms

Tajamul Ashraf[1(✉)], Suhaib Salmani[2], Mohsen Peerzada[2], Ufaq Khan[1], Yutong Xie[1], and Janib Bashir[2]

[1] MBZUAI, Abu Dhabi, UAE
tajamul21.ashraf@gmail.com
[2] National Institute of Technology Srinagar, Srinagar, J&K, India

Abstract. A major challenge in medical imaging is the limited availability of large, well-annotated datasets. In the case of breast cancer detection from mammograms (BCDM), obtaining precise bounding box annotations for regions of interest is costly and labor-intensive. However, large collections of unannotated mammograms are often readily available. Motivated by this observation, we propose a self-supervised fine-tuning framework for BCDM. Traditional object detection models, designed for natural images with abundant object presence, struggle in medical imaging due to limited annotated data. To tackle this challenge, we introduce **MedMask**, a novel self-supervised framework that leverages masked autoencoders (MAE) with vision foundation models (VFMs) in a transformer-based architecture. We propose a customized MAE module that utilizes the transformer's encoder and an auxiliary decoder to mask and reconstruct multi-scale feature maps, enabling efficient learning from limited annotations while capturing domain-specific features. Additionally, we leverage the zero-shot capabilities of VFMs with a proposed expert contrastive knowledge distillation technique to learn better representations. Our approach outperforms the state-of-the-art on the publicly available INBreast and DDSM datasets, achieving significant sensitivity improvements of 22% and 17%, respectively. Also we achieved 27% improvement on RSNA-BSD1K dataset. Code is available at https://github.com/Tajamul21/MedMask.

Keywords: Breast Cancer Detection · Medical Imaging · Mask Autoencoders · Vision Foundational Models

1 Introduction

Breast cancer is among the limited cancers that can be effectively screened. The recent success of vision transformers [9,23,24,41,46] and vision foundation models (VFMs) [2,19,27,36] in medical imaging has significantly accelerated research into the potential for automated triage of mammograms. These models have

T. Ashraf and S. Salmani—Equal contribution.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
T. Zhang et al. (Eds.): Deep-Breath 2025, LNCS 16142, pp. 340–350, 2026.
https://doi.org/10.1007/978-3-032-05559-0_34

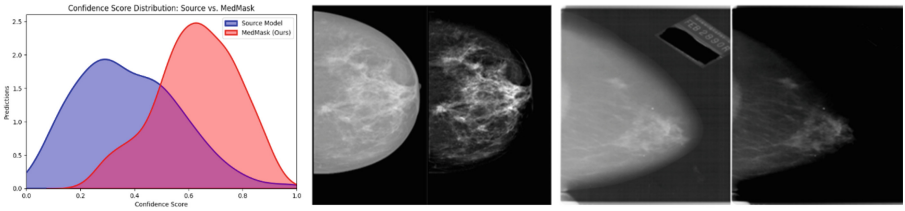


Fig. 1. (a) Our customized `MAE` enhances confidence calibration, leading to more confident and reliable predictions in **MedMask** compared to the Def-DETR predictions (Baseline). (b) We pre-process using Gamma normalization, followed by normalization using the 5th and 99.9th percentile to improve the under-exposed mammograms in `INBreast` (left) and `DDSM` (right) datasets.

already been shown to enhance radiologist's performance in detecting malignancies during screening [26,38,39]. However, the reliability of such networks remains a significant hurdle that must be overcome before these techniques can be adopted in clinical practice. Whereas, automated triage of mammograms requires high sensitivity at a very low false positive rate per image (`FPI`), the current state of the art in breast cancer detection merely provides a sensitivity of 0.685, at an `FPI` of 1% and a sensitivity of 0.021 at an `FPI` of 10%. This work aims to develop a model with high sensitivity at a low `FPI`, making it suitable for triage.

Unlike standard natural images, which typically feature prominent objects, screening mammography focuses on detecting subtle abnormalities. Masses, asymmetries, and microcalcifications are subtle, requiring specialized imaging for detection. We approach this as an object detection problem. Unlike other works [12,33,38,45] formulating the problem as classification, we believe the detection formulation would allow a deep neural network to focus on small but clinically important regions. Note that the detection formulation also enables the explicit demarcation of regions responsible for classifying a mammogram as malignant or benign, making the proposed model highly interpretable–an important objective in computer-aided diagnosis [11,30].

The primary challenge in developing an object detection model for medical imaging is the lack of annotated datasets. While generating manual annotation for any dataset is tedious, it is especially harder for medical images due to the requirement of medical expertise. Here we make an important observation: while manual annotation is costly, the unannotated mammogram is easily available. We propose to fine-tune an object detector in a self-supervised fashion. Unlike traditional Masked Autoencoders (`MAE`) [14] that rely on pretraining on large-scale datasets, we introduce a customized `MAE` module as it better captures spatial semantics critical for medical detection, unlike distillation methods, which focus on global alignment. Our approach randomly masks regions of multi-scale feature maps and reconstructs the missing details using the detection model's encoder and an auxiliary decoder, all while optimizing the detection loss. As a self-supervised task on unlabeled images, this method enables the Transformer encoder to capture better domain-specific patterns from a limited set of ground

truth bounding boxes. Empirically, we observe that MAE enhances confidence calibration, encodes better features, improves performance under limited supervision, and yields larger gains with fewer box annotations, as shown in Fig. 1(a).

Furthermore, since the detection model is sensitive to noise in medical images [39,40], we design a novel expert branch that leverages VFMs to effectively guide the detection model in learning efficient representations. The key insight of this paper is that learned representations from such VFMs can be used to guide a detection model in clinically relevant settings. Thus, we propose contrastive knowledge distillation to allow the VFMs to guide the detector. As shown in the red curve of Fig. 1(a), our method enhances prediction reliability by boosting confidence levels and restoring information corrupted by noise and perturbations.

Contributions of this Work. 1) We propose a novel MedMask framework for training detection transformers using mammograms in self-supervised setting, overcoming the problem of low-quality detections. **2)** We demonstrate that masked autoencoders improve data efficiency in medical imaging and introduce an expert VFMs branch to address local optima. **3)** Our approach significantly surpasses previous methods and sets a new state-of-the-art (SOTA) on three BCDM benchmarks. To showcase the generalizability of our method across diverse settings, we achieve SOTA results on unsupervised domain adaptation tasks in natural settings.

2 Methodology

We propose a three-step training routine, leveraging a large dataset of unannotated mammograms alongside a limited number of bounding box annotations. Figure 2 gives a visual description of our proposed workflow.

2.1 Stage 1 - Baseline Object Detector

Pre-processing: Before inputting the mammograms to our model, we first rescale them from a bit depth of 16 to 8. We observe that mammograms in INBreast and DDSM are underexposed and hence exhibit lower contrast when rescaled uniformly. Thus, we apply gamma normalization with $\gamma = 0.4$ to rescale the mammograms to the $0 - 1$ range and raise pixel values to the power of $1/\gamma$, enhancing their dynamic range. We also normalize the pixel intensities as described below to reduce the effect of extreme values on the dynamic range as shown in Fig. 1(b). We observe that the breast occupies only a portion of the image, while the remaining pixels are mostly black. This causes naive normalization to reduce contrast in important regions. The normalization is given by:

$$I_{\text{norm}} = \frac{I - I_{\text{min}}}{I_{\text{max}} - I_{\text{min}}} \qquad I_{\text{norm}} = \frac{I - I_5}{I_{99.9} - I_5} \tag{1}$$

where I is the original pixel intensity, I_{min} and I_{max} are minimum and maximum intensity values. To avoid outliers and suppress background, we use percentile-based normalization with I_5 and $I_{99.9}$ as the 5th and 99.9th percentile values.

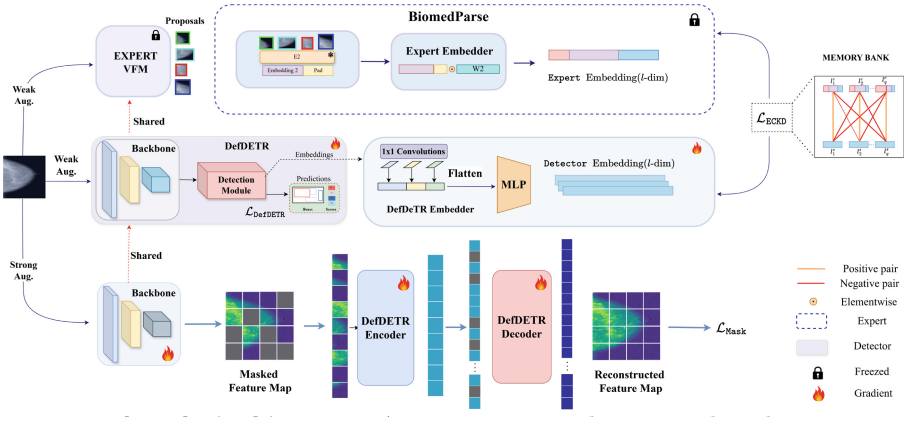


Fig. 2. MedMask Architecture. A mammogram undergoes weak and strong augmentations through a shared backbone. The weak view is also passed to a pretrained expert VFM (BioMedParse) to extract expert features. **E** and **W** represent expert and weak-invariant embeddings, fused for contrastive distillation via $\mathcal{L}_{\text{ECKD}}$. Detection uses DefDETR, while a masked autoencoder (MAE) branch self-supervises on unlabeled data.

Baseline Object Detector: We selected Deformable DETR (`DefDETR`) [46] as our baseline due to its strong generalization. The model processes the resized 640×640 mammograms through a CNN backbone to generate feature maps, which are refined using deformable attention for capturing long-range dependencies. Predicted bounding boxes are assigned objectness scores indicating tumor presence. To improve our pipeline, we replaced the `VGG` [34] backbone with ResNet-101 [15] for better feature learning and differentiation between malignant and benign tumors. The supervised loss $\mathcal{L}_{\texttt{DefDETR}}$ is computed for image x using its corresponding labels y as follows: $\mathcal{L}_{\texttt{DefDETR}} = \mathcal{L}_S^{\text{box}}(x, y) + \mathcal{L}_S^{\text{giou}}(x, y) + \mathcal{L}_S^{\text{cls}}(x, y)$.

2.2 Stage 2 - Masking Module

In the second stage, we develop a mechanism to utilize unannotated mammograms and refine the predictions from Stage 1. We propose a `DefDETR` based Mask Autoencoder (`MAE`) module, acting as a self-supervised adapter to leverage unlabeled mammograms. The multi-scale feature maps from the mammograms will be randomly masked and processed through the `MAE` encoder, with missing features being reconstructed by an auxiliary decoder.

Feature Masking: Unlike the ViT [9] backbone, which processes image patches, the `DefDETR` encoder operates on multi-scale feature maps $\{z_i \in \mathbb{R}^{C_i \times H_i \times W_i}\}_{i=1}^{K}$, where K represents the number of scales, and these feature maps are derived from the backbone output. At each scale, we randomly generate feature masks $\{m_i \in \{0,1\}^{H_i \times W_i}\}_{i=1}^{K}$ with a uniform masking ratio. Deformable attention uses feature map structures to generate reference points, guiding sparse attention matrix by selecting relevant values. As a result, the spatial structure cannot be easily serialized. To address this, we apply zero-masking to the corresponding

value matrix within the deformable attention mechanism [46], in contrast to [14], which only feeds unmasked patches as serialized input to VIT.

Reconstruction: The encoder processes the masked feature maps and further encodes them through deformable attention. We replace the masked regions in the encoder features with a shared mask query (q_m) and feed them into an auxiliary MAE decoder (S_m) to restore the feature map. Following [14], we use an asymmetric and lightweight decoder. As the final output (z_K) from the backbone encapsulates all necessary semantic information, we focus the reconstruction only on this output to speed up convergence and reduce computational load. The MAE decoder's final layer applies a linear projection, matching the output channels to z_K. The loss is computed as the mean squared error (MSE) between the reconstructed and original feature maps, focusing only on masked regions:

$$z_K = S_m(S_e(mask(m, x_t)), q_m) \tag{2}$$

$$\mathcal{L}_{\text{mask}} = \mathcal{L}_{\text{MSE}}(\text{mask}(m_k, z_K), \text{mask}(m_k, \hat{z}_K)) \tag{3}$$

where S_e denotes the encoder of the student model, S_m represents the auxiliary MAE decoder, and q_m is the mask query.

2.3 Stage 3 - VFM Guided Fine Tuning

Experts Branch. We propose a novel framework leveraging Vision Foundational Models (VFMs) through expert distillation (Fig. 2). Assuming VFMs can represent diverse images from datasets like ImageNet [8] or LAION-400M [32], we integrate the pretrained BioMedParse [43] to extract expert embeddings, excelling in tasks like segmentation and detection.

Expert Contrastive Knowledge Distillation Loss (ECKD)**.** Our objective is to transfer the expertise of the expert VFM model to the detection model. The expert model $E(\cdot)$ is frozen, while the detection model is open to gradient updates during the training. Given a mini-batch of images $\{x_1, x_2, \ldots, x_B\}$, for each image $x_i \in \mathcal{X}$, the expert and student models encode it into embeddings E^i_{expert} and E^i_{base}, respectively, as:

$$E^i_{\text{expert}} = \text{Expert}(x_i) \tag{4}$$

$$E^i_{\text{base}} = \mathcal{S}(\mathcal{B}(x_i)) \tag{5}$$

where $\mathcal{B}$ is the student backbone and $\mathcal{S}$ is the projection head.

Our proposed **Expert Contrastive Knowledge Distillation** (ECKD) loss encourages the student embedding E^i_{base} to be similar to the expert embedding E^i_{expert} (positive pair), while ensuring dissimilarity to embeddings E^j_{expert} from other images x_j (negative pairs, $j \neq i$). To address batch bias commonly seen in contrastive learning, we incorporate a memory bank mechanism into the ECKD loss, enabling a larger pool of negative samples without increasing the batch size. Specifically, embeddings from the expert branch are stored in a fixed-size queue,

where the oldest entries are replaced as new mini-batches are processed. This enhances training stability and performance. The ECKD loss is formulated by:

$$\mathcal{L}_{\texttt{ECKD}} = -\log \frac{\texttt{EXP}\left(\left(\overline{E}^{i}_{\text{base}}\right)^{\top} \overline{E}^{i}_{\text{expert}}/\tau\right)}{\sum_{q=1}^{Q} \texttt{EXP}\left(\left(\overline{E}^{i}_{\text{base}}\right)^{\top} \overline{E}^{q}_{\text{expert}}/\tau\right)} \tag{6}$$

where $\overline{E}$ are ℓ_2-normalized embeddings, τ is the temperature, B is the batch size, and $Q > B$ is the memory bank size.

2.4 Overall Objective and Training Strategy

We jointly train the MAE module and detection loss (DefDETR) instead of adopting a pretraining-finetuning approach, as pretraining could cause the model to overfit to the reconstruction task due to the small dataset and distinct medical domain characteristics. Note that we apply MAE only to unannotated images to improve detection on the unlabeled target population. Applying MAE to a small number of annotated images may bias domain features and degrade performance. Using unannotated images helps the MAE encoder better understand the domain, enhancing performance with limited high-quality bounding boxes. We first train the model using the supervised loss $\mathcal{L}_{\texttt{DefDETR}}$ and the masking loss $\mathcal{L}_{\text{mask}}$ (Eq. (3)) to obtain an enhanced initialization θ_{mask} for VFM guided fine-tuning as shown:

$$\mathcal{L}_{\texttt{MAE}} = \mathcal{L}_{\texttt{DefDETR}} + \lambda_{\text{mask}} \mathcal{L}_{\text{mask}} \tag{7}$$

where $\mathcal{L}_{\texttt{DefDETR}}$ is the bounding box loss defined in Subsect. 2.1, and λ_{mask} is the weighting coefficient for $\mathcal{L}_{\text{mask}}$. The overall objective after the expert phase is as follows:

$$\mathcal{L} = \mathcal{L}_{\texttt{MAE}} + \lambda_{\texttt{ECKD}} \mathcal{L}_{\texttt{ECKD}} \tag{8}$$

where $\mathcal{L}_{\texttt{MAE}}$ and $\mathcal{L}_{\texttt{ECKD}}$ are defined in Eqs. (7) and (6) respectively. As training progresses, we gradually decay both coefficients i.e. λ_{mask} and $\lambda_{\texttt{ECKD}}$ to prevent the encoder from overfitting, detailed analysis is provided in Subsect. 3.2.

3 Experiments

Datasets. The datasets used in our experiments are as follows: **(1)** DDSM [20]: A publicly available dataset with 1324 mammography images, including 573 cases of malignancy. **(2)** INBreast [26]: Contains 200 images from 115 patients, with 41 malignant and 159 benign cases. **(3)** RSNA-BSD1K [5]: The original RSNA dataset [5] includes 54,706 mammograms from 8,000 patients, with 1,000 malignancies. The curated RSNA-BSD1K subset consists of 1,000 mammograms, with 200 malignant cases and bounding box annotations. We used 70/20/10 (train/val/test) split for each dataset. Only malignant lesions were used for evaluation to match clinical triage focus.

Table 1. Comparison with state-of-the-art OD techniques.

Datasets	Model Name	Venue	R@0.05	R@0.1	R@0.3	R@0.5	R@1.0	R@2.0
INBreast [26]	FRCNN [29]	NIPS'15	0.38	0.38	0.44	0.44	0.44	0.55
	PISA [3]	CVPR'20	0.28	0.41	0.41	0.57	0.63	0.70
	CenterNet2 [44]	ArXiv'21	0.31	0.38	0.45	0.56	0.61	0.73
	DETR [4]	ECCV'20	0.05	0.09	0.14	0.20	0.26	0.32
	DefDETR [46]	ICLR'21	0.33	0.38	0.66	0.72	0.77	0.88
	DN-DEF [21]	CVPR'22	0.11	0.18	0.26	0.33	0.41	0.50
	Dab-DETR [23]	ICLR'22	0.09	0.18	0.22	0.30	0.35	0.39
	CO-DETR [47]	ICCV'23	0.11	0.15	0.19	0.23	0.23	0.28
	YOLO-V5 [16]	UTX'23	0.22	0.35	0.47	0.50	0.62	0.68
	DINO [13]	ICLR'23	0.31	0.52	0.68	0.77	0.81	0.81
	MedMask (Ours)		**0.66**	**0.77**	**0.88**	**0.88**	**0.94**	**1.00**
DDSM [20]	FRCNN [29]	NIPS'15	0.16	0.21	0.30	0.33	0.47	0.48
	PISA [3]	CVPR'20	0.10	0.15	0.28	0.39	0.44	0.57
	CenterNet2 [44]	ArXiv'21	0.27	0.38	0.43	0.44	0.52	0.60
	DETR [4]	ECCV'20	0.12	0.15	0.22	0.32	0.32	0.39
	DefDETR [46]	ICLR'21	0.29	0.47	0.69	0.76	0.85	**0.90**
	DN-DEF [21]	CVPR'22	**0.35**	0.41	0.52	0.61	0.61	0.78
	Dab-DETR [23]	ICLR'22	0.11	0.23	0.35	0.47	0.55	0.63
	CO-DETR [47]	ICCV'23	0.21	0.28	0.31	0.31	0.47	0.51
	YOLO-V8 [16]	UTX'23	0.56	0.56	0.56	0.56	0.60	0.63
	DINO [13]	ICLR'23	0.16	0.20	0.28	0.30	0.33	0.38
	MedMask (Ours)		0.32	**0.49**	**0.72**	**0.79**	**0.85**	0.88
RSNA-BSD1K [5]	FRCNN [29]	NIPS'15	0.06	0.12	0.38	0.38	0.44	0.44
	PISA [3]	CVPR'20	0.18	0.23	0.35	0.44	0.52	0.59
	CenterNet2 [44]	ArXiv'21	0.01	0.09	0.11	0.15	0.19	0.21
	DETR [4]	ECCV'20	0.08	0.11	0.18	0.25	0.29	0.32
	DefDETR [46]	ICLR'21	0.11	0.29	0.70	0.75	0.90	0.90
	DN-DEF [21]	CVPR'22	0.25	0.29	0.33	0.41	0.29	0.71
	Dab-DETR [23]	ICLR'22	0.01	0.15	0.28	0.28	0.37	0.44
	CO-DETR [47]	ICCV'23	0.21	0.28	0.41	0.44	0.58	0.63
	YOLO-V5 [16]	UTX'23	0.32	0.49	0.69	0.69	0.81	0.89
	DINO [13]	ICLR'23	0.31	0.31	0.63	0.73	0.91	0.91
	MedMask (Ours)		**0.41**	**0.52**	**0.88**	**0.93**	**0.95**	**0.95**

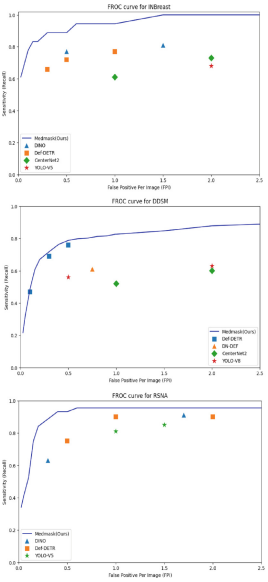


Fig. 3. FROC curves

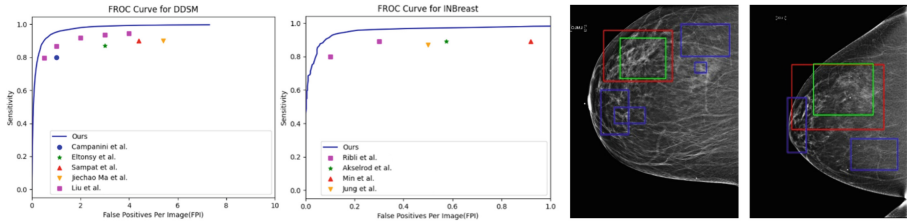
Evaluation Metric. We report our results using Free-Response Receiver Operating Characteristic (FROC) curves [10], which graphically represent sensitivity/recall at various false positives per image (FPI). Following [28], a prediction is true positive if its predicted box center lies within the ground-truth box.

Implementation Details. We use a 2-layer asymmetric decoder [14] in the MAE framework with an initial mask ratio of 0.8. The model is optimized with Adam [18], learning rate 2×10^{-5}, and batch size 16. Data augmentation includes random horizontal flips, Gaussian blurring, and random contrast adjustments.

3.1 Comparison with SOTA Object Detectors

Natural Image Techniques. We fine-tune various baseline object detection models, originally proposed for natural image datasets, on BCDM datasets and compare the results. The results are tabulated in Table 1. From these results, we observe that DefDETR performs well in majority of the cases across all datasets. Hence, we choose DefDETR as our baseline and apply the proposed 3 stage pipeline on it. Figure 3 shows visual results by SOTA object detectors on the benchmark datasets. Note that Table 1 also shows corresponding results using our proposed model. Hence, the table also serves as comparison of our proposed model with the SOTA object detectors proposed for natural images.

Breast Cancer Detection Techniques. We compare the results of our proposed model with the results reported by existing works on breast cancer detection. We present the results using FROC curves as shown in Fig. 4(a) and (b) on the predefined test sets on INBreast and DDSM datasets respectively. We show that our model outperforms all existing breast cancer detection SOTA models.



(a) (b) (c)

Fig. 4. (a) and (b) shows comparison with Breast Cancer `SOTA` detection models [1,17,25,30], (c) shows qualitative analysis with red, yellow, and green boxes as `GT`, `DefDETR` [46], and `MedMask` respectively. (Color figure online)

Table 2. A comprehensive ablation study of the `MAE` hyperparameters (*left*), proposed modules (*middle*), and the generalizability of **MedMask** in natural settings (*right*).

Category	Parameter	mAP
Mask ratio	0.5	47.1
	0.8	**53.1**
	0.9	46.8
Coefficient decay	Linear	48.2
	No decay	47.9
	Hard	**53.1**

Module	Baseline	Retrain	VFM	MAE	R@0.4
	✓				0.72
	✓			✓	0.85
	✓		✓		0.79
	✓	✓	✓		0.80
	✓	✓		✓	0.87
	✓	✓	✓	✓	**0.93**

Model	Venue	mAP
AT [22]	CVPR'22	47.4
PT [6]	ICML'22	42.2
LGCL [35]	BMVC'24	46.7
O2net [37]	ISPRS'20	46.8
MIC [42]	ECCV'22	43.4
MedMask (Ours)	-	**53.1**

3.2 Ablation Study

Quantitative and Qualitative Analysis. The impact of the proposed modules is summarized in Table 2(middle), where we observe the following: 1) The inclusion of the `MAE` branch (line 2) leads to a significant improvement over the source-only model, highlighting `MAE` as a data-efficient learner. 2) The `MAE`, expert, and retraining modules each result in substantial performance gains (lines 3–5) when used individually, confirming their individual effectiveness. 3) Combining the expert branch with `MAE` (line 7) leads to further improvements, demonstrating that our design leverages the strengths of both modules. Qualitative results on the `INBreast` dataset are provided in Fig. 4(c).

Analysis of `MAE` Module and Genralizability. Table 2(left) indicates that a higher mask ratio enhances the effectiveness of the self-supervision. We also examine `MAE` coefficient λ_{mask} decay strategies: "no decay" keeps λ_{mask} fixed, "linear" decays linearly, and "hard" drops the `MAE` branch after a set epoch. "No decay" underperforms as the `MAE` branch benefits less from low-confidence predictions and needs reduced influence in later stages. Our contributions are generalizable beyond medical imaging and can benefit natural image settings also. Therefore, we evaluate our model against semi-supervised techniques using Cityscapes [7] to Foggy Cityscapes [31] as shown in Table 2(right).

4 Conclusion

The limited adoption of deep neural networks across diverse geographies, data distributions, and imaging protocols hinders their clinical adoption for breast cancer detection. This paper introduces a new semi-supervised framework, **MedMask**, to tackle this issue. Our customized masked autoencoder module enables the model to capture domain-specific characteristics and learn from limited annotations, while the expert branch helps avoid local optima influenced by noise in medical imaging. Experimental results across four diverse datasets from various domains show that our approach outperforms `SOTA`, particularly in recall, with low false positives per image. We hope our work will enhance the clinical adoption of automated breast cancer detection models by improving generalization using our framework.

References

1. Akselrod-Ballin, A., et al.: Deep learning for automatic detection of abnormal findings in breast mammography. In: DLMIA 2017, Held in Conjunction with MICCAI (2017)
2. Awais, M., et al.: Foundational models defining a new era in vision: a survey and outlook. In: TPAMI (2024)
3. Cao, Y., Chen, K., Loy, C.C., Lin, D.: Prime sample attention in object detection. In: Proceedings of the IEEE/CVF CVPR, June 2020
4. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: ECCV, pp. 213–229. Springer (2020)
5. Carr, C., et al.: Rsna screening mammography breast cancer detection (2022)
6. Chen, M., et al.: Learning domain adaptive object detection with probabilistic teacher. arXiv:2206.06293 (2022)
7. Cordts, M., et al.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE CVPR, pp. 3213–3223 (2016)
8. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: a large-scale hierarchical image database. In: 2009 IEEE CVPR, pp. 248–255. IEEE (2009)
9. Dosovitskiy, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale. In: ICLR (2021)
10. Egan, J.P., Greenberg, G.Z., Schulman, A.I.: Operating characteristics, signal detectability, and the method of free response. In: JASA (1961)
11. Févry, T., P, J., Wu, N., Kim, S., Moy, L., Cho, K., Geras, K.J.: Improving localization-based approaches for breast cancer classification. In: MIDL (2019)
12. Geras, K.J., et al.: High-resolution breast cancer screening with multi-view deep convolutional neural networks. arXiv preprint arXiv:1703.07047 (2017)
13. Zhang, H., et al.: DINO: DETR with improved denoising anchor boxes for end-to-end object detection. In: ICLR (2023)
14. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: CVPR (2022)
15. He, K., et al.: Deep residual learning for image recognition. In: CVPR (2016)
16. Jocher, G., Chaurasia, A., Qiu, J.: Ultralytics yolov8 (2023). https://github.com/ultralytics/ultralytics

17. Jung, H., et al.: Detection of masses in mammograms using a one-stage object detector based on a deep convolutional neural network. PLoS ONE **13**(9), e0203355 (2018)
18. K, D.P., Ba, J.: Adam: a method for stochastic optimization. In: ICLR (2015)
19. Kirillov, A., et al.: Segment anything. In: Proceedings of the IEEE/CVF ICCV, pp. 4015–4026 (2023)
20. Lee, R.S., Gimenez, F., Hoogi, A., Miyake, K.K., Gorovoy, M., Rubin, D.L.: A curated mammography data set for use in cad and diagnosis research. Sci. Data (2017)
21. Li, F., Zhang, H., Liu, S., Guo, J., Ni, L.M., Zhang, L.: Dn-detr: accelerate detr training by introducing query denoising. In: CVPR (2022)
22. Li, Y.J., et al.: Cross-domain adaptive teacher for object detection (2022)
23. Liu, S., et al.: Dab-detr: dynamic anchor boxes are better queries for detr. In: ICLR (2022)
24. Meng, D., et al.: Conditional detr for fast training convergence. In: ICCV (2021)
25. Min, H., Chandra, S.S., Crozier, S., Bradley, A.P.: Multi-scale sifting for mammographic mass detection and segmentation. BPEE **5**(2), 025022 (2019)
26. Moreira, I.C., Amaral, I., Domingues, I., Cardoso, A., Cardoso, M.J., Cardoso, J.S.: Inbreast: toward a full-field digital mammographic database. Acad. Radiol. (2012)
27. Oquab, M., et al.: Dinov2: learning robust visual features without supervision. In: TMLR (2024)
28. Rangarajan, K., et al.: Deep learning for detection of iso-dense, obscure masses in mammographically dense breasts. Eur. Radiol. (2023)
29. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: NuerIPS (2015)
30. Ribli, D., Horváth, A., Unger, Z., Pollner, P., Csabai, I.: Detecting and classifying lesions in mammograms with deep learning. Sci. Rep. **8**(1), 4165 (2018)
31. Sakaridis, C., Dai, D., Van Gool, L.: Semantic foggy scene understanding with synthetic data. Int. J. Comput. Vision **126**, 973–992 (2018)
32. Schuhmann, C., et al.: Laion-400 m: open dataset of clip-filtered 400 million image-text pairs. In: NeurIPS (2021)
33. Shu, X., Zhang, L., Wang, Z., Lv, Q., Yi, Z.: Deep neural networks with region-based pooling structures for mammographic image classification. In: IEEE TMI (2020)
34. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: ICLR (2015)
35. Triantafyllidou, D., Parisot, S., Leonardis, A., McDonagh, S.: Improving object detection via local-global contrastive learning. In: BMVC (2024)
36. Wang, H., et al.: Sam-clip: merging vision foundation models towards semantic and spatial understanding. In: NeurIPS (2023)
37. Wei, H., Zhang, Y., C, Z., Li, H., Wang, H., Sun, X.: Oriented objects as pairs of middle lines. ISPRS J. Photogram. Remote Sens. **169** (2020)
38. Wu, Nan, e.a.: Breast density classification with deep convolutional neural networks. In: 2018 IEEE ICASSP, pp. 6682–6686. IEEE (2018)
39. Wu, N., et al.: Deep neural networks improve radiologists' performance in breast cancer screening. IEEE TMI **39**(4), 1184–1194 (2019)
40. Xie, Y., Yin, Y., Li, Q., Wang, Y.: Deep mutual distillation for semi-supervised medical image segmentation. In: MICCAI (2023)
41. Yang, J., Li, C., Dai, X., Gao, J.: Focal modulation networks. In: NeurIPS (2022)

42. Yu, J., et al.: Mttrans: cross-domain object detection with mean teacher transformer. In: ECCV (2022)
43. Zhao, T., et al.: A foundation model for joint segmentation, detection and recognition of biomedical objects across nine modalities. Nat. Methods **22**(1), 166–176 (2024). https://doi.org/10.1038/s41592-024-02499-w, http://dx.doi.org/10.1038/s41592-024-02499-w
44. Zhou, X., Koltun, V., Krähenbühl, P.: Probabilistic two-stage detection. arXiv preprint arXiv:2103.07461 (2021)
45. Zhu, W., Lou, Q., Vang, Y.S., Xie, X.: Deep multi-instance networks with sparse label assignment for whole mammogram classification. In: MICCAI (2017)
46. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: deformable transformers for end-to-end object detection. In: ICLR (2021)
47. Zong, Z., Song, G., Liu, Y.: Detrs with collaborative hybrid assignments training. In: Proceedings of the IEEE/CVF ICCV, pp. 6748–6758 (2023)

Longitudinal Mammogram Exam-Based Breast Cancer Diagnosis Models: Vulnerability to Adversarial Attacks

Zhengbo Zhou[1], Degan Hao[1], Dooman Arefan[2], Margarita Zuley[2], Jules Sumkin[2], Lars Grimm[5], James Joshi[3], and Shandong Wu[1,2,4(✉)]

[1] Intelligent Systems Program, Pittsburgh, USA
[2] Department of Radiology, 203 Lothrop Street, Pittsburgh, PA 15213, USA
[3] Department of Informatics and Networked Systems, Pittsburgh, USA
[4] Biomedical Informatics and Bioengineering, University of Pittsburgh, Pittsburgh, PA, USA
wus3@upmc.edu
[5] Department of Radiology, Duke University, Durham, NC, USA

Abstract. In breast cancer detection and diagnosis, the longitudinal analysis of mammogram images is crucial. Contemporary models excel in detecting temporal imaging feature changes, thus enhancing the learning process over sequential imaging exams. Yet, the resilience of these longitudinal models against adversarial attacks remains underexplored. In this paper, we propose a novel blackbox attack approach that capitalizes on the feature-level relationship between two sequential mammogram exams of a longitudinal model, guided by both cross-entropy loss and distance metric learning, to achieve significant attack efficacy as implemented using attack transferring in a black-box attacking manner. We perform experiments on a cohort of 590 breast cancer patients (each has two sequential mammogram exams) in a case-control setting. Results show that our proposed method surpasses several state-of-the-art adversarial attacks in fooling the diagnosis models to give opposite outputs. Our method remains effective even if the model is trained by employing existing defense mechanisms for adversarial attacks.

Keywords: Breast cancer · Adversarial attack · Longitudinal model · Diagnosis

1 Introduction

Mammography-based AI is at the forefront of medical AI research, and many AI products are being translated for clinical deployment. The cybersecurity of such models is therefore a paramount to ensure AI integrity. Although medical IT systems are relatively closed, cybersecurity incidents still occur, quite often, through infiltration of the IT systems [14] by external threat agents or by insider attacks [16]. For example, hospitals have had to pay ransom when their data are

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
T. Zhang et al. (Eds.): Deep-Breath 2025, LNCS 16142, pp. 351–361, 2026.
https://doi.org/10.1007/978-3-032-05559-0_35

compromised through ransomware attacks [9]. As per [8], 94 % of U.S. hospitals are affected by healthcare data breaches, showing high vulnerability and risks to cybersecurity attacks. As elucidated in [6], adversarial attacks can occur because of malicious intents or activities through ransomware attacks, insurance fraud, manipulation of clinical trial data, etc.

In clinical radiology, the established practice of comparing prior data with current data serves as a cornerstone for radiologists to perform lesion detection and diagnosis. The emergence of deep learning models has refined this process by integrating temporal sequential data, such as using multiple mammograms, magnetic resonance imaging, computed tomography, etc., as model inputs. Recent evidences indicate that these longitudinal models outperform those relying solely on a single time-point input. Cui et al. [3] show that applying convolutional neural networks to follow-up MRI scans can effectively diagnose Alzheimer's disease. Dadsetan et al. [4] propose a novel deep learning model on longitudinal mammograms showing superior effects compared to a model using a single exam. Lee et al. [12] Leverage a Transformer decoder to incorporate prior images, showing improved performance on breast cancer risk prediction.

In the biomedical domain, longitudinal models using sequential imaging scans are gaining popularity, where the models are made more effective by utilizing spatiotemporal relationships of longitudinal data. Adversarial attacks on models using individual time-point scans have received considerable attention [22]. However, there has been little or no exploration into attacks on longitudinal models.

In this work, we focus on the task of diagnosing breast cancer using a Transformer decoder architecture [19] that uses sequential mammogram exams. We examine the susceptibility of this longitudinal model to adversarial attacks. We propose a novel black-box attack method that combines cross-entropy loss (to guide adversarial samples across the decision boundary) and distance metric learning (to modify the relationship between sequential imaging exams), aiming to fool the diagnosis models to generate opposite outputs. We utilize the adversarial attack transferability to achieve this, and show that, even when adversarial training is employed to enhance the model's robustness, our proposed attack can still effectively degrade the model's performance. Our method also outperforms several other compared methods. Our study highlights a significant vulnerability of longitudinal models to adversarial attacks, thus highlighting the critical need to enhance such models' resilience against adversarial attacks.

The contributions of our work can be summarized as: (1) We studied a novel topic of investigating adversarial attacks on longitudinal imaging-based diagnostic models; (2) We evaluated the method of integrating cross-entropy loss with distance metric learning to implement the attack, which exhibited stronger attaching effects in leading to misclassification of breast cancer, compared to several state-of-the-art attack techniques; and (3) We showed that in distance metric learning, medical knowledge can be leveraged in selecting effective adversarial samples aiming to fool a diagnosis model.

2 Related Work

Adversarial Attack Methods: Szegedy et al. [18] show that classifiers may confidently make incorrect predictions when subjected to imperceptible perturbations. Kurakin et al. [10] show that classifiers remain vulnerable to adversarial samples even in the physical world. Goodfellow et al. [7] propose FGSM, a simple and efficient adversarial sample-generating method. In addition, Basic Iterative Method (BIM) [11] is proposed as an extension of the Fast Gradient Sign Method (FGSM) and aims to generate stronger adversarial samples by iteratively applying FGSM with smaller perturbation steps. Benefiting from the iteration strategy, MI-FGSM is proposed by adding momentum through iterations [5]. The Projected Gradient Descent (PGD) attack [15] is another popular and powerful attack method. On the other hand, the optimization-based C&W attack [1] treats adversarial sample generation as an optimization problem.

Attack Transferability: The vulnerability of models to adversarial samples generated by other models is known as the transferability of adversarial attacks [18]. Kurakin et al. [10] conduct a study on adversarial samples within the context of ImageNet; they show that BIM's multi-step approach is less transferable than the single-step FGSM. Zhou et al. [23] show that enhancing the transferability of the BIM can be achieved by maximizing the distance between natural images and their adversarial counterparts within the intermediate feature maps, coupled with the addition of regularization. The transferability of a single-input model to longitudinal models is an interesting approach, but has not yet been studied in the literature.

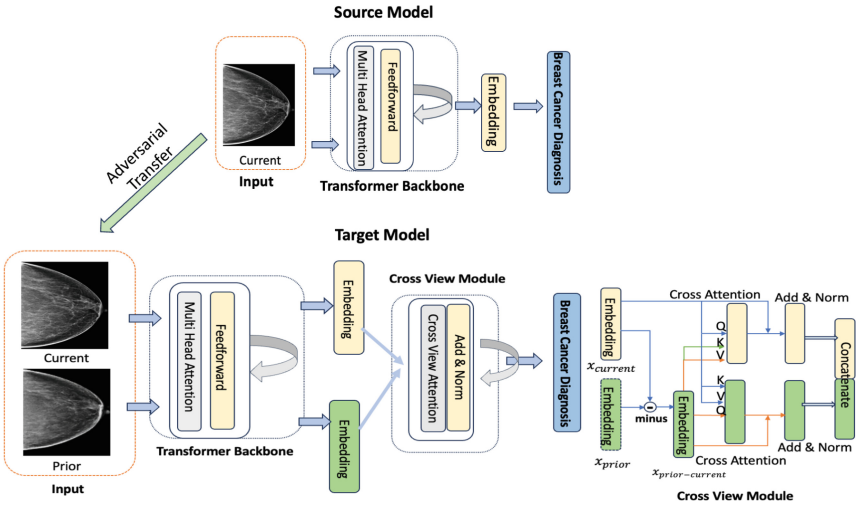


Fig. 1. Adversarial attacks transferred from a source model (a single input, Current) to a target model (two sequential input, Prior and Current).

3 Methods

3.1 Task and Model Architecture

Task. In this study, we design and implement our novel method on the task of diagnosing breast cancer on mammograms (i.e., classification of cancer cases vs. normal controls). The diagnosis can be based on a single time-point exam or multiple longitudinal exams that leverage temporal information to improve diagnosis accuracy. Thus, for the propose attack, we design two models: Source model (using a single time-point mammogram exam, denoted as Current) and Target model (using two sequential exams, denoted as Prior and Current, respectively, which are taken at two different time points). The Source model is useful when a patient does not have Priors available. It should be pointed out that the Prior exams are normal/healthy mammograms that are stored as historical data of a patient in the electronic health records. The use of normal Priors in longitudinal models conforms to radiologists' clinical practice to improve lesion detection and diagnosis, in addition to using the Current exam.

Model Architecture. As illustrated in Fig. 1, in the Source model, features extracted by the backbone model are utilized for diagnosis, which also serve as the basis for generating adversarial samples to attack the target longitudinal model i.e., (the attack transferring). For the Target longitudinal model, the imaging features are derived from both the Prior and Current exams, referred to as x_{prior} and x_{current}, respectively. The inputs to the cross-view module are current exam x_{current} and the changes between the prior and current exam, i.e., $x_{\text{prior-current}} = x_{\text{prior}} - x_{\text{current}}$. Here $x_{\text{prior-current}}$ aims to capture and emphasize the changes between the two time points. Subsequently, both feature vectors are fed to the cross-view module. In this module, we use multi-head attention mechanisms to identify and stress the information in the subtracted feature ($x_{\text{prior-current}}$) that is relevant to the current feature (x_{current}). The core of multi-head attention is a scaled dot-product attention [19] mechanism that computes attention scores between queries and keys, which are then used to linearly combine the associated values. In our setting, we reshape the embedded feature maps for the target phase (x_{current} or $x_{\text{prior-current}}$) into a query matrix and reshape the feature maps for the source phase into a key matrix. The outcome of this attention process produces $x_{\text{current-attention}}$ and $x_{\text{(prior-current)-attention}}$. To form new attention-enhanced feature vectors, we integrate these representations with the original vectors, resulting in $x_{\text{(prior-current)-cross}} = x_{\text{prior-current}} + x_{\text{(prior-current)-attention}}$ and $x_{\text{current-cross}} = x_{\text{current}} + x_{\text{current-attention}}$. These refined feature vectors are then concatenated, serving for the ultimate prediction.

3.2 Adversarial Attack

Rationale: Our study aims to attack the Target model that makes diagnoses using the temporal relationships/changes of the Prior and Current exams. In our black-box attack scenario, attackers do not have access to the model architecture and parameters of the Target model, particularly the mechanisms in which

the longitudinal features are used. Attackers craft adversarial samples using the Source model, following the scheme of "attack transferring", to perform attacks on the Target model. This makes sense because attackers do not need to or may not have a longitudinal model to generate adversarial samples; instead, a Source model based on a single time-point input is much more accessible and easily available to use. Here, the attacks aim to manipulate the Current exam without considering the Prior exam, as the cancer diagnosis is primarily based on the Current exam (recall that the Prior exams are all normal/healthy and they serve as comparisons to Current exams to improve cancer detection).

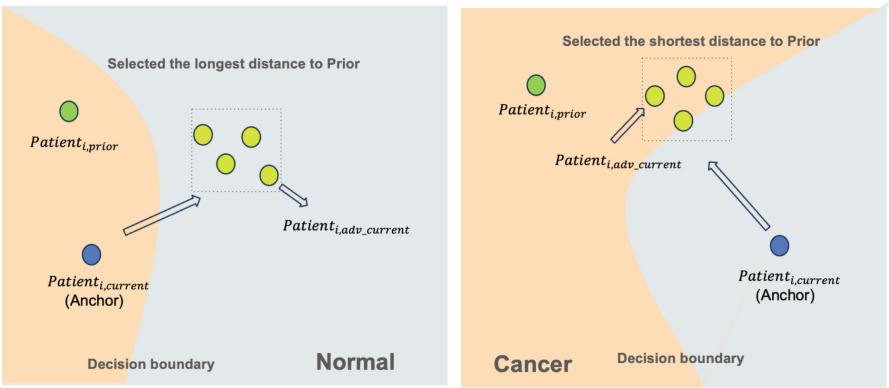


Fig. 2. Illustration of knowledge-guided selection of samples for adversarial Current: to select the one with the shortest (longest) distance to Prior for Cancer (Control) cases. Light orange represents Normal decision space, while gray represents Cancer decision space. (Color figure online)

Adversarial Sample Generation. Let $f(s)$ denote an arbitrary deep neural network which takes s ($s \in \mathbb{R}^n$) as input and outputs the probability of classes y ($y \in \mathbb{R}^m$). We first define an adversarial sample fooling the model $f(s)$ for a chosen p-norm and noise parameter ϵ as follows:

$$\tilde{s} = \text{argmax}_{\|\tilde{s}-s\|_p \leq \epsilon} l(\tilde{s}, t) \tag{1}$$

where t and $l(\cdot, \cdot)$ denote the label of x and the loss function used to train the model, respectively. In our experiments, we investigate the effects of using the cross-entropy loss and distance metric learning for attacking the Target model. To optimize Eq. (1), the Iterative Fast Gradient Sign Method (I-FGSM) is employed. Initially, the input x is scaled to the range into $[-1, 1]$ and we set $\tilde{x}_0 = x$. Then, we compute the gradients of the loss function with respect to input x. Following this, the adversarial samples undergo several iterations of updates. In each step, we apply the sign function to the calculated gradients and then clip the updated adversarial samples to the range $[-1, 1]$ to maintain them as valid images. At every iteration, the generated samples are designed to

cross the decision boundary, directed by the gradient data, and are retained for subsequent sampling. Ultimately, adversarial samples are generated by adding the pixel differences, scaled by ϵ, between the most recently updated adversarial samples and the original input x.

Knowledge-guided Adversarial Sample Selection Method. As illustrated in Fig. 2, the I-FGSM method can generate multiple adversarial samples aimed at traversing the decision boundary. We propose novel criteria to select the most impacting adversarial samples using distance metric learning. This criteria is designed based on the knowledge of: (1) the longitudinal model relies on the relationship of Prior (always normal) and Current (could be cancer or normal) exams to make a diagnosis, and (2) the Euclidean distance between Prior and Current is smaller for a normal patient (because of less variation across two normal exams), but larger for a cancer patient (because of larger variation from normal transitioning to cancer development). Specifically, the criteria differ between Normal and Cancer cases: in Cancer cases, we select the adversarial sample $Patient_{i,adv_current}$ that is closest from $Patient_{i,prior}$, while in Control cases, we choose the sample that is furthest to $Patient_{i,prior}$. For Cancer cases, the objective is to coax the model into misclassifying the adversarial sample as Normal by leveraging the proximity to a $Patient_{i,prior}$ image labeled as normal. This proximity strategy enhances the likelihood that the longitudinal model, which relies on temporal relationships for classification, erroneously assesses the sample as Normal. Conversely, for Normal cases, the intent behind selecting adversarial samples with the greatest distance from the Prior image is to challenge the model into a false diagnosis of Cancer. Given the substantial separation between the adversarial sample and a Prior image, the longitudinal model is more inclined to misclassify these as belonging to Normal cases. This novel adversarial sample selection approach takes advantage of the longitudinal model's dependencies on the relationship between $Patient_{i,prior}$ and $Patient_{i,current}$ exams.

4 Experiments

4.1 Study Cohort

This study received institutional review board approval, and we used a dataset of 590 subjects in a case-control study setting, with 293 breast cancer cases and 297 breast cancer-free controls (i.e., normal/negative). Each subject has a Current mammogram exam and a Prior exam taken at approximately 1 year apart. All diagnosis outcomes are biopsy-proven and based on the Current exams. All Prior exams are normal/negative. For the cancer cases, we exclusively utilized the biopsied breast for. For the controls, either the right or left breast was randomly chosen for a subject to avoid modeling from shortcut learning. To ensure uniformity in the orientation of a breast, we applied a horizontal flip for a right-sided breast to appear like a left-sided breast. We specifically opted for using the craniocaudal view of the mammogram images.

4.2 Implementation Details

The dimensions of the input mammogram images were standardized to 350×400 pixels to maintain consistency across the subjects. We employ a Swin Transformer [13] pretrained on ImageNet as the backbone for feature extraction in both the Source and Target models. We also evaluate the effects of implementing the Source model with a VGG [17] model as the backbone architecture, whereas the Target model remains unchanged. To prevent overfitting for both the Source and Target model training, we incorporate data augmentation during the training phase, including flipping and image rotation. The model is trained over 30 epochs, with learning rates of 5e–5 and 1e–5 identified as the optimal parameter values through experiments. For generating adversarial attacks, the iteration number was set to 15 for all iterative-based attack methods, and the perturbation size was set to 0.01 for all attacks. We perform parameter robustness analysis experiments on these two parameters.

We compare the effects of our method to several methods, including C&W, FGSM, I-FGSM, MI-FGSM, and PGD. Also, we compare ours to two distance-guided new methods using distance loss on FGSM and I-FGSM, respectively, as described by Eq. (2), where we denote x as input data, t as label, L(x) as the intermediate feature map:

$$\text{loss}_{\text{distance}}(x', t) = \begin{cases} L(x') - L(x_{\text{prior}}) & \text{if } t = 1 \\ L(x_{\text{prior}}) - L(x') & \text{if } t = 0 \end{cases} \tag{2}$$

In addition, we have implemented two more novel methods that use distance metric loss to alter the relationship between Prior and adversarial Current. Specifically, we use the distance loss in Eq. (2) as a penalty term aiming to more efficiently guide the search directions. We refer to this regularization method as 'Distance Reg.' (Eq. (3)) as another way to combine the cross entropy loss and distance metric learning. We compare our proposed knowledge-guided method to this alternative method (λ is experimentally determined as 0.05).

$$\tilde{s} = \text{argmax}_{\|\tilde{s}-s\|_p \leq \epsilon} l(\tilde{s}, t) + \lambda \cdot \text{loss}_{\text{distance}}(s', t) \tag{3}$$

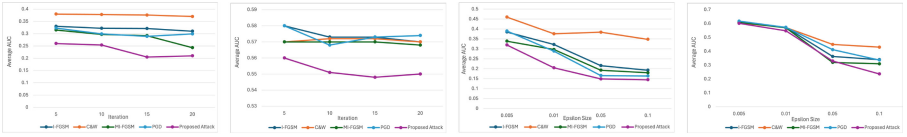
We utilize a patient-wise 5-fold cross-validation and Area Under the ROC Curve (AUC) reporting Mean AUC ± Standard Deviation (Std) of the various methods. We also assess the model's performance after incorporating adversarial training as a defence method, where the model underwent retraining with a combination of clean data and adversarial samples generated by BIM with a perturbation size of 0.01 and a batch size of 32. All computational tasks are carried out on an NVIDIA TESLA V100 GPU, provided by our local supercomputing facility.

5 Results

Table 1 presents comprehensive attack performance on Swin Transformer and VGG source backbones, with Target models using Swin Transformer. Target

Table 1. Performance of the evaluated adversarial attacks when the source model is a Transformer or VGG. Metrics are reported for the corresponding Source model, Target model without and with adversarial training (mean AUC ± SD).

Attack Method	Source = Swin Transformer			Source = VGG		
	Source	Target	Target (Adv.)	Source	Target	Target (Adv.)
No Adv. Attack	0.670 ± 0.02	0.704 ± 0.05	0.685 ± 0.03	0.660 ± 0.05	0.704 ± 0.05	0.685 ± 0.03
FGSM [7]	0.194 ± 0.05	0.410 ± 0.04	0.584 ± 0.04	0.161 ± 0.06	0.546 ± 0.05	0.597 ± 0.04
Dist.-guided FGSM	0.403 ± 0.06	0.531 ± 0.04	0.634 ± 0.03	0.365 ± 0.04	0.545 ± 0.05	0.615 ± 0.03
Dist. Reg. FGSM	0.292 ± 0.08	0.452 ± 0.05	0.606 ± 0.03	0.292 ± 0.08	0.452 ± 0.05	0.606 ± 0.03
I-FGSM [11]	0.000 ± 0.00	0.322 ± 0.05	0.573 ± 0.04	0.000 ± 0.00	0.450 ± 0.08	0.595 ± 0.04
Dist.-guided I-FGSM	0.039 ± 0.01	0.430 ± 0.03	0.612 ± 0.04	0.000 ± 0.00	0.515 ± 0.08	0.603 ± 0.04
Dist. Reg. I-FGSM	0.000 ± 0.00	0.286 ± 0.05	0.573 ± 0.03	0.000 ± 0.00	0.447 ± 0.05	0.596 ± 0.03
C&W [1]	0.000 ± 0.00	0.376 ± 0.06	0.572 ± 0.04	0.000 ± 0.00	0.502 ± 0.07	0.580 ± 0.03
MI-FGSM [5]	0.000 ± 0.00	0.297 ± 0.05	0.570 ± 0.04	0.000 ± 0.00	0.446 ± 0.08	0.591 ± 0.04
PGD [15]	0.000 ± 0.00	0.289 ± 0.06	0.573 ± 0.03	0.000 ± 0.00	0.461 ± 0.11	0.596 ± 0.02
AutoAttack [2]	0.000 ± 0.00	0.308 ± 0.04	0.585 ± 0.03	0.000 ± 0.00	0.514 ± 0.07	0.585 ± 0.02
GI-FGSM [21]	0.000 ± 0.00	0.286 ± 0.04	0.587 ± 0.03	0.000 ± 0.00	0.475 ± 0.08	0.617 ± 0.01
PC-I-FGSM [20]	0.000 ± 0.00	0.291 ± 0.05	0.584 ± 0.02	0.000 ± 0.00	0.530 ± 0.07	0.626 ± 0.01
GRA [24]	0.000 ± 0.00	0.303 ± 0.04	0.574 ± 0.04	0.000 ± 0.00	0.464 ± 0.11	0.614 ± 0.01
Proposed Attack	0.000 ± 0.00	**0.205** ± 0.04	**0.548** ± 0.04	0.000 ± 0.00	**0.418** ± 0.10	**0.577** ± 0.02



(a) (b) (c) (d)

Fig. 3. Parameter robustness analysis. Target model performance is shown with respect to a range of values of the iteration number (A, B) and perturbation size (C, D). Adversarial training was applied in B and D.

models initially show higher baseline AUCs than Source models, highlighting the utility of prior exams. Adversarial attacks consistently reduce AUCs across both backbones to 0.205–0.531 range with a Swin-Transformer source and to 0.418–0.546 with a VGG source, effectively deceiving the models. Our proposed attack consistently achieves the lowest AUCs (0.205 for Swin Transformer source; 0.418 for VGG source), outperforming all other methods and demonstrating robust effectiveness across architectures. The adversarial training method involves incorporating adversarial samples into the training set to enhance model robustness through retraining [10]. Adversarial training (Target (Adv.)) improves Target model resilience, raising AUCs to 0.548–0.634 range with a Swin-Transformer source and to 0.577–0.615 with a VGG source compared to non-trained models, though clean data performance slightly drops. Attacks still degrade performance despite training. Our proposed attack remains most

effective, yielding the lowest AUCs (0.548 for Swin Transformer source; 0.577 for VGG source) even after adversarial training. Analysis also shows cross-entropy, PGD, and distance-guided methods independently degrade performance. Distance Regularization approaches surpass distance-guided ones but are less effective than our proposed knowledge-guided sample selection, which is a superior attack strategy. Figure 3 illustrates our method outperforming I-FGSM, C&W, MI-FGSM, and PGD across various parameter values, with or without adversarial training. AUCs predictably decrease as epsilon (perturbation strength) increases.

6 Conclusion

In this study, we have delved into the medical imaging AI model robustness against adversarial attacks on longitudinal models, with a particular focus on breast cancer diagnosis. Our research contribution is novel because studies on adversarial attacks to longitudinal models are rare, yet such models are gaining popularity in medical applications. We have proposed a novel adversarial attack that combines cross-entropy loss and knowledge-guided distance metric learning, showing superior effects fooling the diagnosis model, and outperforming several state-of-the-art methods. Our method remains effective even after incorporating the defense methods of adversarial training. Future work includes further evaluation of different model structures and developing effective defense methods. Our study highlights the importance of adversarially robust medical diagnosis models towards delivering safe AI to patient care.

Acknowledgments. This work was supported in part by a National Institutes of Health (NIH) Other Transaction research contract #1OT2OD037972-01; a National Science Foundation (NSF) grant (CICI: SIVD; #2115082); grant 1R01EB032896 as part of the NSF/NIH Smart Health and Biomedical Research in the Era of Artificial Intelligence and Advanced Data Science Program; an Amazon Machine Learning Research Award; and the University of Pittsburgh Momentum Funds (Scaling Grant) for the Pittsburgh Center for AI Innovation in Medical Imaging. This work used Bridges-2 at the Pittsburgh Supercomputing Center through allocation [MED200006] from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by NSF grants #2138259, #2138286, #2138307, #2137603, and #2138296. This research was also supported in computing resources by the University of Pittsburgh Center for Research Computing and Data (RRID: SCR_022735) through the resources provided by the H2P cluster, which is supported by NSF award OAC-2117681. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing official policies, either expressed or implied, of the NIH or NSF.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Carlini, N., Wagner, D.: Towards evaluating the robustness of neural networks. In: 2017 IEEE Symposium on Security and Privacy (SP), pp. 39–57. IEEE (2017)
2. Croce, F., Hein, M.: Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In: International Conference on Machine Learning, pp. 2206–2216. PMLR (2020)
3. Cui, R., Liu, M., Initiative, A.D.N., et al.: Rnn-based longitudinal analysis for diagnosis of alzheimer's disease. Comput. Med. Imaging Graph. **73**, 1–10 (2019)
4. Dadsetan, S., Arefan, D., Berg, W.A., Zuley, M.L., Sumkin, J.H., Wu, S.: Deep learning of longitudinal mammogram examinations for breast cancer risk prediction. Pattern Recogn. **132**, 108919 (2022)
5. Dong, Y., et al.: Boosting adversarial attacks with momentum. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 9185–9193 (2018)
6. Finlayson, S.G., Bowers, J.D., Ito, J., Zittrain, J.L., Beam, A.L., Kohane, I.S.: Adversarial attacks on medical machine learning. Science **363**, 1287–1289 (2019)
7. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014)
8. HIPAA Journal: Healthcare data breach statistics (nd)
9. Jr., B.L.: Hospital ceo forced to pay hackers in bitcoin now teaches others how to deal with ransomware. In: CNBC, April 2018
10. Kurakin, A., Goodfellow, I., Bengio, S.: Adversarial machine learning at scale. arXiv preprint arXiv:1611.01236 (2016)
11. Kurakin, A., Goodfellow, I.J., Bengio, S.: Adversarial examples in the physical world. In: Artificial Intelligence Safety and Security, pp. 99–112. Chapman and Hall/CRC (2018)
12. Lee, H., Kim, J., Park, E., Kim, M., Kim, T., Kooi, T.: Enhancing breast cancer risk prediction by incorporating prior images. arXiv preprint arXiv:2303.15699 (2023)
13. Liu, Z., et al.: Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022 (2021)
14. Lyngaas, S.: Cyberattack is a factor in illinois hospital's closure. CNN, June 2023
15. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083 (2017)
16. Mirsky, Y., Mahler, T., Shelef, I., Elovici, Y.: Ct-gan: malicious tampering of 3d medical imagery using deep learning. 2019 (2019)
17. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
18. Szegedy, C., et al.: Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199 (2013)
19. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
20. Wan, C., Huang, F.: Adversarial attack based on prediction-correction. arXiv preprint arXiv:2306.01809 (2023)
21. Wang, J., et al.: Boosting the transferability of adversarial attacks with global momentum initialization. arXiv preprint arXiv:2211.11236 (2022)
22. Zhou, Q., et al.: A machine and human reader study on ai diagnosis model safety under attacks of adversarial images. Nat. Commun. **12**(1), 7281 (2021)

23. Zhou, W., et al.: Transferable adversarial perturbations. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 452–467 (2018)
24. Zhu, H., Ren, Y., Sui, X., Yang, L., Jiang, W.: Boosting adversarial transferability via gradient relevance attack. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4741–4750 (2023)

Author Index

A
Adalid-Llansa, Laia 165
Adewole, Maruf 289
Ahmed, Faisal 31
Anazodo, Udunna C. 289
Arefan, Dooman 351
Ashraf, Tajamul 340
Asiimwe, Rosta 289
Atukunda, Patience 289
Awwad, Hadeel 248

B
Bago-Horvath, Zsuzsanna 329
Baralis, Elena 52
Bashir, Janib 340
Bayer, Siming 258
Bean, Gregory 155
Bernal Rodriguez, Alba 320
Bernathova, Maria 329
Betancourt, Diana Spence 289
Bhat, Sheethal 258
Bonci, Eduard 176, 186, 268
Boonen, Pieter 123
Bounias, Dimitrios 310
Bridge, Christopher P. 63
Brussaard, Carola 123

C
Cardona, Julian 226
Cardoso, Jaime S. 41, 176, 186, 268, 279
Cardoso, Maria J. 41, 176, 186, 268, 279
Casas, Josep R. 165
Castro, André C. 237
Catanese, Alessandro 226
Ceranka, Jakub 123
Cerquitelli, Tania 52
Chen, Han 91
Chen, Yaqian 101, 196
Chen, Yuwen 196
Cleveland, Mason C. 63
Cochet, Alexandre 134
Cruz, Helena 279
Cuba, Julian 226
Curran, Kathleen M. 206
Czipczer, Vanda 1

D
Dako, Farouk 289
Das, Chandisha 216
Dashtbozorg, Behdad 21
Daye, Dania 63
del Riego, Javier 226
Diaz, Oliver 226
Domingo-Ferrer, Josep 31
Dorca, Eduard 165
Dorfner, Felix J. 63
Dorster, Manon A. 63

E
Elbarbary, Karim 258
Elbatel, Marawan 300
Emegoakor, Adaobi 289
Engelmann, Justin 320
Ertl, Alexandra 310
Eze, Franca 289

F
Falcetta, Daniele 52
Farooq, Aiman 216
Fatade, Abiodun 289
Faye, Ibrahima 113
Ferenczi, Lehel 1
Fernández, Pedro L. 165
Ferreira, Pedro 41, 268
Filho, Arlindo R. G. 237

G
Gago, Lucas 320
Gbadamosi, Yewande 289
Geldof, Freija 21
Gonalves, Tiago 41

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
T. Zhang et al. (Eds.): Deep-Breath 2025, LNCS 16142, pp. 363–365, 2026.
https://doi.org/10.1007/978-3-032-05559-0

Gonçalves, Tiago 268
Grimm, Lars J. 196
Grimm, Lars 351
Gu, Hanxue 101, 196
Guelen, Melisa S. 63

H
Haddi, Zouhair 31
Hamdi, Moustapha 123
Hao, Degan 351
Helbich, Thomas 329

I
Ibarra, Sebastian 226
Infante, Jonathan 226
Iorumbur, Aondona M. 289
Itakura, Haruka 155

J
Jatho, Alfred 289
Jimenez-Martin, Lauren 165
Joshi, James 351

K
Kächele, Jessica 310
Kalaiwo, Chinasa 289
Kamran, Mohammed 329
Kang, Solha 81
Kapila, Ayush 123
Kawooya, Michael 289
Khalid, Adnan 113
Khan, Ufaq 340
Kim, Albert E. 63
Kim, Eugene 81
Konz, Nicholas 101

L
Lalande, Alain 113, 134
Lamtenzan, Diego 123
Langs, Georg 329
Lee, Jeong Hoon 155
Lekadir, Karim 226
Leon, Nataly 226
Li, Qihang 101, 196
Lippenszky, Levente 1

M
Maier, Andreas 258
Maier-Hein, Klaus 310
Makhoul, Peter 289
Maksudov, Bulat 206
Malumba, Richard 289
Mann, Ritse 21
Marciano', Vincenzo 52
Martel, Anne L. 91
Martí, Robert 248
Marti, Robert 300
Mata, Daniel 165
Mavioso, Carlos 268
Mazurowski, Maciej A. 101, 196
Meriaudeau, Fabrice 113, 134
Mes, Rebecca 21
Mileo, Alessandra 206
Mishra, Deepak 216
Montenegro, Helena 176, 186
Montoya-del-Angel, Ricardo 300
Muhammad, Abbas Rabiu 289
Mursil, Muhammad 113
Musah, Toufiq 72
Musinguzi, Dennis 289

N
Nakka, Krishna Kanth 11
Nascimento, Jacinto C. 145
Nnamani, Amaka 289

O
Osuala, Richard 226
Ozbulak, Utku 81

P
Panambur, Adarsh Bhandary 258
Pastor, Eliana 52
Patel, Jay 63
Payan, Neree 134
Peerzada, Mohsen 340
Petit, Ana 165
Pinto, Gonçalo 279
Poeta, Eleonora 52
Pons, Laura 165
Puig, Domenec 113

R
Ramos-Oliver, Irma 165
Rashwan, Hatem A. 113
Raymond, Confidence 289
Remeserio, Beatriz 320
Rodrigues, Dayane 237

Ruers, Theo 21
Rusu, Mirabela 155

S
Salazar-Huayna, Lourdes 165
Salmani, Suhaib 340
Sánchez, David 31
Santiago, Carlos 145
Santos, Joana 186
Schouten, Daan 155
Schwarz, Lea 300
Seeböck, Philipp 329
Singer, Christian F. 329
Soler-Monsó, Teresa 165
Sumkin, Jules 351

T
Tan, Tao 21
Tareke, Tewele W. 134
Teixeira, Fábio 176
Teixeira, Luís F. 279
Thiran, Jean-Philippe 63
Timko, Edina 1
Toyobo, Oluyemisi 289

U
Umoren, Charity 289

V
van Karnenbeek, Lennard M. 21
Vandemeulebroucke, Jef 123
Vankerschaver, Joris 81
Varga, Raoul 329
Vargas, Luisa 52
Velasco Rebolledo, Jacinto 320
Vilanova, Joan C. 248
Vilaplana, Verónica 165
Vilardell, Felip 165

W
Wijkhuizen, Mark 21
Wu, Shandong 351

X
Xie, Yutong 340

Y
Yang, Hongxu 1
Yang, Jichen 196

Z
Zhou, Zhengbo 351
Zolfagharnasab, Mohammad Hossein 41, 268, 279
Zuley, Margarita 351
Zuluaga, Maria A. 52