ODELIA

Deliverable D2.3

PUBLIC MRI DATASET INCLUDING 500 EXAMINATIONS

Lead beneficiary	CAM
Author(s)	Dr. Nicholas Payne, Gustav Müller-Franzes
Reviewer(s)	Prof. Dr. Daniel Truhn (UKA)
Dissemination level	PU
Type	DATA
Delivery date	June 30, 2025

**Funded by
the European Union**

TABLE OF CONTENTS

DISCLAIMER

Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency (HADEA). Neither the European Union nor the granting authority can be held responsible for them.

EXECUTIVE SUMMARY

A comprehensive dataset comprising 741 breast MRI examinations has been curated collaboratively by six ODELIA consortium partners (CAM, MHA, RSH, RUMC, UKA, and UMCU), each contributing between 31 and 250 examinations. Each examination has been annotated at the breast level as either "no lesion," "benign lesion," or "malignant lesion." The dataset has been detailed in a pre-print manuscript [https://arxiv.org/abs/2506.00474] and serves as the basis for a publicly accessible AI model development challenge. The dataset is split into training (408 examinations), validation (103 examinations), and test (230 examinations) subsets. The training and validation sets were publicly released in June 2025, with the test set withheld until the challenge concludes in September 2025.

INTRODUCTION

A major barrier to developing robust and generalisable AI models is the limited availability of large, diverse, publicly accessible breast MRI datasets. Existing publicly available datasets typically originate from single-centre studies, are extensively curated, and predominantly include malignant cases only. Such datasets do not adequately represent the clinical heterogeneity encountered in real-world scenarios and limit the training of AI models capable of accurately distinguishing malignant, benign, and non-lesion findings.

To address these limitations, we set out to curate a large, multi-centre breast MRI dataset collected from six ODELIA centres. The resulting dataset encompasses various scanner manufacturers, clinical settings, and acquisition protocols, capturing the variability inherent to real-world clinical practice. It includes not only confirmed malignant lesions but also benign and control/unremarkable cases, making it the largest and most diverse publicly available dataset of its kind.

DATASET CREATION

741 Breast MRI examinations were collected from the following six European medical centers between December 2006 and May 2024:

- CAM: Cambridge University Hospitals, Cambridge, UK
- MHA: Mitera Hospital, Athens, Greece
- RSH: Ribera Hospital, Valencia, Spain
- RUMC: Radboud University Medical Center, Nijmegen, Netherlands
- UKA: University Hospital Aachen, Aachen, Germany
- UMCU: University Medical Center Utrecht, Utrecht, Netherlands

Each center contributed between 31 and 250 studies and imaging protocols adhered to institution-specific clinical standards, ensuring diverse data representation.

Expert radiologists at each center classified lesions in both the left and right breast based on histopathological or 2-year follow-up information (ground truth). Initially, lesions were categorized as:

- **No lesion**: No contrast-enhancing lesion is visible
- **Benign lesion**: A contrast-enhancing lesion is visible but confirmed benign via biopsy or two-year follow-up.
- **Malignant lesion (DCIS)**: A lesion of type ductal carcinoma in situ (DCIS)
- **Malignant lesion (Invasive)**: A malignant invasive lesion
- **Malignant lesion (Unknown)**: A malignant lesion of unknown specific type

All MRI scans were initially acquired in DICOM format and subsequently converted to NIfTI format. A standardized naming and processing scheme was developed, extracting metadata from each DICOM

file and categorizing images into dynamic T1-weighted (T1w) and T2-weighted (T2w) sequences. T1w images were differentiated into pre-contrast and multiple post-contrast phases. A subtraction image (post-contrast minus pre-contrast) was computed to enhance lesion visibility.

For many machine learning applications, a more unified image data format is required. For this purpose, the T1w sequence was resampled to a resolution of 0.7 x 0.7 x 3.0 mm, and the T2w sequence was resampled to the T1w sequence. Subsequently, all examinations were separated into left and right breast regions by dividing the images at the centre. Furthermore, using a threshold to separate background and foreground, breasts were center cropped or padded to 256 x 256 x 32 voxels.

The final dataset was divided into three distinct subsets: training (408), validation (103), and testing (230), ensuring appropriate evaluation and benchmarking of AI algorithms.

PUBLIC CHALLNEGE

The curated dataset is being utilized for a public AI challenge hosted on the Grand Challenge platform. The challenge was officially launched in May 2025, with training and validation subsets released in June 2025. Currently, 39 teams have registered to participate.

The withheld test set will serve as the final benchmark for submitted AI models. Models will be evaluated based on the following metrics:

- Area Under the Receiver Operating Characteristic Curve (AUC)
- Specificity at 90% sensitivity
- Sensitivity at 90% specificity

Final rankings will be determined by averaging raw ranks of these metrics. Ties will be resolved by prioritizing the highest AUC score.

This challenge has been accepted for presentation at the MICCAI 2025 conference as part of the Deep-Breath workshop, further validating the significance and quality of this initiative.

CONCLUSION

The ODELIA consortium has successfully created a comprehensive, highly diverse breast MRI dataset from multiple European institutions. More than 500 examinations have been publicly released to facilitate AI model training and validation. The remaining test dataset will be publicly released following the completion of the challenge, enhancing the development and validation of robust, clinically relevant AI solutions.