README BreastDCEDL dataset

BreastDCEDL

Code: https://github.com/naomifridman/BreastDCEDL

Authors: Naomi Fridman, Itamar Barnea, Tomer Fridman and Bubby Solway.

BreastDCEDL is a curated collection of pretreatment 3D dynamic contrast-enhanced MRI (DCE-MRI) scans from 2,070 breast cancer patients, assembled into a deep learning–ready dataset. It integrates data from three major clinical trials: I-SPY2 (n = 982), I-SPY1 (n = 172), and Duke (n = 916). The dataset, originally sourced from The Cancer Imaging Archive (TCIA), includes:

- 3D raw MRI scans converted to NIfTI format
- Corresponding 3D tumor binary segmentation masks
- Clinical and demographic metadata, including pCR, HER2, HR, age, and race

BreastDCEDL Data Download

I-SPY1: The complete I-SPY1 dataset is available for direct download from Zenodo.

Duke: The full Duke cohort can be accessed via The Cancer Imaging Archive (TCIA). Conversion to NIfTI format can be performed using the provided code in the GitHub repository. A minimized version—containing three (n_z,256×256) tumor-centered scans per patient—is available on Zenodo.

I-SPY2: The full I-SPY2 dataset is available on TCIA and can be converted to NIfTI using the GitHub code. A pre-converted NIfTI version will be made available on TCIA in the near future.

Zendoo Data Organization of I SPY-1 and DUKE

duke.zip → Contains duke_dce/ + BrestDCEDL_duke_cropped.csv

spy1.zip → Contains spy1_dce/ + BreastDCEDL_spy1_metadata.csv + spy1_mask/

models.zip → Contains Breastdcedl_pcr_vit_model_weights.pth

Top-level files:

- README_BreastDCEDL_dataset.pdf
- LICENSE
- overview_breastdcedl.png

The Zenodo upload includes a full version of the **I-SPY1** cohort and a **minimized version of DUKE**, where each patient has only three selected DCE-MRI scans:

- **Scan 0** – pre-contrast
- **Scan 1** – first post-contrast
- **Final scan** – the last available post-contrast scan (typically scan 2–4)

Directories and Files:

- **duke_dce/** — Contains 3 DCE-MRI 3D NIfTI volumes per patient (cropped around tumor).
- **spy1_dce/** — Contains 3 to 5 DCE-MRI 3D NIfTI volumes per patient from the I-SPY1 cohort.
- **`BrestDCEDL_duke_cropped_metadata.csv`** — Metadata for DUKE patients, including cropping parameters and bounding box coordinates.
- **BrestDCEDL_spy1_metadata.csv** — Metadata for I-SPY1 patients with clinical labels and imaging features.
- **Breastdcedl_pcr_vit_model_weights.pth** — Pretrained Vision Transformer (ViT) weights for binary pCR prediction.

Each metadata file includes patient-level labels (pCR, HR, HER2), scan identifiers, and preprocessing info.

In BrestDCEDL_duke_cropped_metadata.csv, the following fields document the original and cropped tumor bounding boxes:

- **Original bounding box:** org_sraw, org_eraw, org_scol, org_ecol, org_mask_start, org_mask_end, org_n_z, org_n_xy
- **Cropped bounding box:** crop_sraw, crop_eraw, crop_scol, crop_ecol, crop_mask_start, crop_mask_end
- **Final selected ROI in cropped image:** sraw, eraw, scol, ecol, mask_start, mask_end

These values allow accurate mapping between full and cropped space for reproducible tumor localization and spatial analysis.

This structure supports seamless integration into machine learning pipelines for breast cancer outcome prediction.

Dataset Details

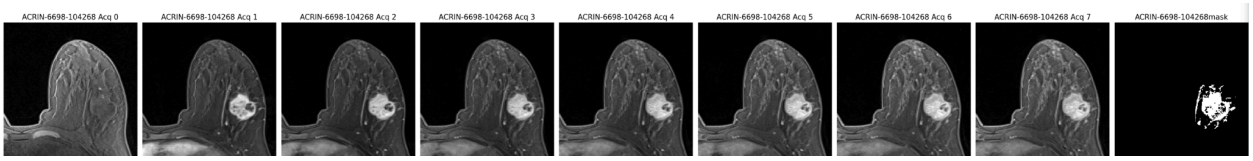
🧪 I-SPY2 Dataset

The I-SPY2 trial (Li et al., 2022; Newitt et al., 2021) provides DCE-MRI scans for 982 patients acquired from 2010 to 2016 across over 22 clinical centers using a standardized imaging protocol.

- Target cohort: Women with high-risk, locally advanced breast cancer
- Clinical data: pCR, HR, HER2, MammaPrint (MP) scores, type of neoadjuvant therapy, age, and race

🖼️ Imaging Details:

- Each MRI scan includes 3 to 12 time points (mostly 7)
- Radiologists selected 3 time points for tumor segmentation: typically scans 0 (pre-contrast), 2 (early post-contrast), and 5 or 6 (late post-contrast). These selections are provided in the metadata under `pre`, `post_early`, and `post_late`.

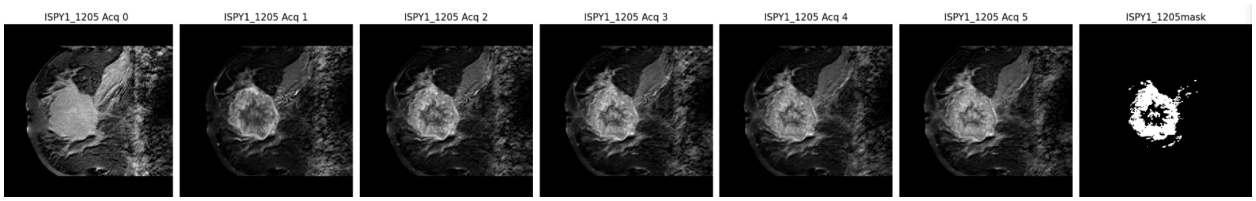


🧪 I-SPY1 Dataset

The I-SPY1 dataset is a predecessor to I-SPY2 and contains similar imaging and clinical information, with slightly fewer patients and minor differences in acquisition protocols.

- Patients: 172 with 3–5 usable DCE scans

- Clinical data: pCR, HR, HER2, and other core biomarkers



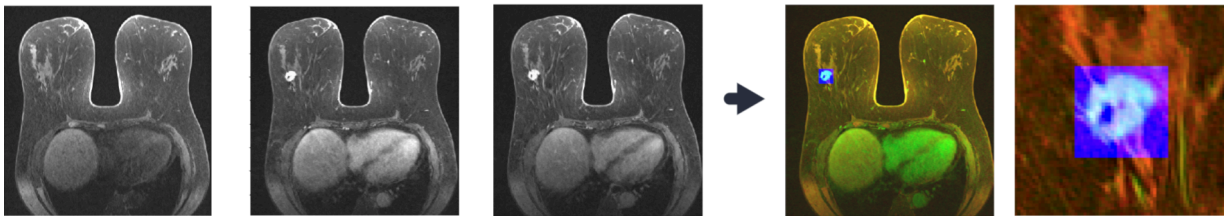
🧪 Duke Dataset

The Duke Breast Cancer Dataset consists of 916 patients with biopsy-confirmed invasive breast cancer, collected between 2000 and 2014.

- Only 288 patients (31%) received neoadjuvant chemotherapy (NAC) and have annotated pCR values.
- The rest underwent surgery first, followed by adjuvant therapy, and are not included in pCR analysis.
- DCE-MRI scans include one pre-contrast and 2–4 post-contrast acquisitions, spaced 1–2 minutes apart.

🖼️ Data Processing Notes:

- Bounding box annotations of the largest tumor are provided.
- No full tumor segmentation masks are available for Duke.



Benchmark Prediction Tasks

The dataset provides a standardized benchmark for three central classification tasks in breast cancer MRI:

- Pathological Complete Response (pCR): A binary classification task predicting treatment response based on pretreatment imaging. Approximately 32.2% of patients (n = 317) achieved pCR, offering a moderately balanced class distribution. *pCR (pathologic complete response) refers to the complete disappearance of all invasive cancer cells in the breast and lymph nodes following neoadjuvant therapy and is considered a strong surrogate for favorable long-term prognosis.*
- Hormone Receptor (HR) Status: Classification of HR positivity (present in 54.5% of cases, n = 537) directly from imaging, assessing the link between MRI features and receptor expression.
- HER2 Status: Prediction of HER2 expression (positive in 24.8%, n = 244) from imaging data, enabling evaluation of MRI-based biomarker inference.

Split	pCR N	pCR+	pCR−	HR N	HR+	HR−	HER-	HER+	HER2−
Train	1099	322	777	1529	987	542	1528	345	1183
Val	177	53	124	268	163	101	269	58	210
Test	176	53	123	271	173	98	268	56	213
Total	1452	428	1024	2068	1327	741	2065	459	1606

Note: `pCR N` refers to the number of patients with non-missing pCR labels; similarly, `HR N` and `HER2 N` indicate the number of patients with available HR and HER2 status, respectively. Class distributions are shown for each split.

DCE-MRI Clinical Background

Dynamic Contrast-Enhanced MRI (DCE-MRI) is a 3D imaging technique that captures a sequence of scans before and after the injection of a contrast agent (typically gadolinium). The contrast enhances visibility of blood vessels and tissue perfusion, allowing observation of how the agent accumulates and clears from tissues over time.

Tumors exhibit characteristic enhancement patterns: malignant lesions often enhance quickly and wash out, while benign lesions typically enhance more slowly or steadily. Radiologists assess these patterns by reviewing two or three key time points—commonly the pre-contrast image and one or two post-contrast phases (e.g., the 2nd, 3rd, or 4th scan in the series). This helps them distinguish between benign and malignant lesions and informs treatment decisions.

These enhancement dynamics are critical both for clinical evaluation and for machine learning models that aim to predict malignancy, treatment response, or other tumor characteristics.

Source

All datasets were originally acquired from:

- The Cancer Imaging Archive (TCIA)
- Monticciolo et al., 2018, *Journal of the American College of Radiology (JACR)*
- ClinicalTrials.gov - I-SPY2 (NCT01042379)