Data in Brief 57 (2024) 110995

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

LA-Breast: A Latin American multiparametric breast DCE-MRI dataset with benign and malignant annotations

Rubén D. Fonnegra[a,b,*], Carlos Mera[a], Gloria M. Díaz[a], Liliana Hernández[c]

[a] Instituto Tecnológico Metropolitano, Colombia
[b] Institución Universitaria Pascual Bravo, Colombia
[c] Ayudas Diagnósticas SURA, Colombia

ARTICLE INFO

Article history:
Received 12 June 2024
Revised 22 September 2024
Accepted 27 September 2024
Available online 5 October 2024

Dataset link: LA-Breast DCE-MRI Dataset (Original data)

Keywords:
DCE-MRI
Annotated
Benign lesions
Malignant lesions

ABSTRACT

This dataset was collected from single-institution retrospective examinations from Latin American patients. Patients were anonymized and contains 15 DCE-MR imaging sequences: pre-contrast T1 fat saturated Dynamic (d0), five postcontrast T1 fat saturated dynamics (d1 to d5), T1 with no fat saturation (t1), T2 with no fat saturation (t2), the apparent diffusion coefficient image (ADC) and the diffusion image (Diff), and five post-contrast phase sequences (f1 to f5). All image sequences were obtained and stored using the standard DICOM 3.0 and converted to TIFF format to reduce required storage space. Besides, all images were obtained using multiple 1.5T scanners and all contrast agents were gadolinium-based with dosages between 0.014 and 0.016 l/mol under different acquisition conditions as they were obtained retrospectively. Each patient data was filtered according to their available clinical finding, were benign and malignant lesions were prioritized to ensure at least one relevant clinical finding across all images. Thereupon, the data contains balanced train, test and validation sets in terms of benign/malignant lesions, as well as non-dense/dense tissue. Additionally, annotations per image are provided with lesion

* Corresponding author.
E-mail address: ruben.fonnegra@pascualbravo.edu.co (R.D. Fonnegra).
Social media: @RubenFonnegra (R.D. Fonnegra)

https://doi.org/10.1016/j.dib.2024.110995

location (x and y coordinates) BIRADS and tissue density. This data can be used for multiple purposes, including image synthesis, image characterization, lesion classification, tissue segmentation, among others.

Specifications Table

Subject	Health and medical sciences
Specific subject area	Medical imaging
Data format	Raw, Analyzed, Filtered
Type of data	Images
Data collection	Images were selected from retrospective studies in the archive of the Instituto de Alta Tecnología Médica (IATM), Medellín Colombia. Each study was anonymized and contains 15 imaging sequences: pre-contrast T1 fat saturated Dynamic (d0), five postcontrast T1 fat saturated dynamics (d1 to d5), T1 with no fat saturation (t1), T2 with no fat saturation (t2), the apparent diffusion coefficient image (ADC) and the diffusion image (Diff), and five post-contrast phase sequences (f1 to f5). All image sequences were obtained and stored using the standard DICOM 3.0 and converted to TIFF format to reduce required storage space. Besides, all studies were acquired using multiple 1.5T scanners and all contrast agents were gadolinium-based with dosages between 0.014 and 0.016 l/mol as they were obtained retrospectively.
Data source location	Instituto de Alta Tecnología Médica (IATM)
Data accessibility	Repository name: LA-Breast DCE-MRI Dataset Data identification number: 10.17632/8rzyn3ng9c.1 Direct URL to data: https://data.mendeley.com/datasets/8rzyn3ng9c/1
Related research article	[Fonnegra et. al] [2], April). Early-to-Late Prediction of DCE-MRI Contrast-Enhanced Images in Using Generative Adversarial Networks. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)* (pp. 1-5). IEEE.

1. Value of the Data

- The data includes annotated malignant and benign lesions from DCE-MRI patients according to BIRADS and ACR, with such variability being uncommon in this imaging modality.
- The data contains balanced train, test and validation sets in terms of benign/malignant lesions, as well as non-dense/dense tissue.
- The data was collected retrospectively, which ensures multiple acquisition and variability conditions.
- Data contains multiple MRI image sequences to ensure completeness, including 15 different modalities for analysis.
- All studies focus on Latin American patients, a demographic for which there is limited data available for analysis in DCE-MRI.
- This data can be used for multiple purposes, such image synthesis, image characterization, lesion classification, among others.

2. Background

Contrast-enhanced magnetic resonance imaging (DCE-MRI) has proven to be the most sensitive technique for detecting breast cancer. This method is particularly effective in identifying small tumors and those in dense breast tissues, which are often obscured by other imaging modalities such as mammography and ultrasound [1]. The DCE-MRI acquisition protocol includes

Table 1
Acquisition parameters for the studies. Studies were selected from a retrospective dataset to preserve diversity but ensure homogeneity.

	Axial T1W VISTA	DWI B0 and B800	T1 SPIR THRIVE1	Axial T2 VISTA
Repetition time	7.5	7112	6.8	7010
Echo time	4.6	70	3.3	80
Field of view (mm)	280 × 368 × 180	300 × 400 × 198	300 × 337 × 156	280 × 370 × 180
Acquisition time	3:44	2:36	6:07	3:16
Num of pre-contrast sequences	1	1	1	1
Num of post-contrast sequences	-	-	5 (60.1s/seq.)	-
Post-processing sequences	-	ADC maps	Phase (subtraction)	-

a series of anatomical images, followed by five or more images dynamically captured after the application of a gadolinium-based contrast agent [3]. The high sensitivity of this imaging technique derives from analyzing the dynamic behavior of the contrast agent within the imaging findings, specifically assessing whether it persists or washes out [4]. However, the use of DCE-MRI is limited by its high costs, lengthy acquisition and analysis times, and the side effects associated with the contrast agent, which have been widely documented [5].

This dataset was developed to explore alternatives to using contrast agents, employing computer vision and artificial intelligence techniques. Potential avenues include developing detection and classification models for findings in anatomical images or synthesizing images that simulate the response to contrast agents from anatomical images alone. The authors of this work are currently investigating these innovative approaches.

3. Data Description

The dataset is splitted into multiple subfolders. The root path contains folders with the name of each sequence (d0, d1, d2….) for ease to use. A total of 15 folders with images can be found. Additionally, three more folders can be found within each image sequence directory, containing the train, test and validation samples. Images are anonymized and stored in TIFF format to preserve the raw image data and reduce the required storage space. Images are named according to the following convention: *"PatientID_ROIID_OriginalDCMfilename.tiff"* (e.g., Mri_1_R1_IM-1682-0111.tiff). An illustration of the structure tree of the repository is shown in Fig. 1. Finally, the root directory contains a folder named metadata which contains CSV files containing additional information per image, such the lesion location, name of file, region identifier and location. For more details about the content of each file, please refer to Table 1.

4. Experimental Design, Materials and Methods

The images selected for this work correspond to the dataset compiled from the project grant RC 740, Minciencias. Images were selected from retrospective studies in the archive of the Instituto de Alta Tecnología Médica (IATM), in Medellín, Colombia. Thereupon, all images in the dataset correspond to Latin American patients. As studies were carefully selected, they were selected according to the acquisition parameters, with the aim to preserve diversity but ensure homogeneity among them. Acquisition parameters considered are shown in Table 1. Each study was anonymized following the convention "MRI_N" where N represents the integer number of the patient (between 1 and 200). Each study contains 15 imaging sequences: pre-contrast T1 fat saturated Dynamic (d0), five postcontrast T1 fat saturated dynamics (d1 to d5), T1 with no fat saturation (t1), T2 with no fat saturation (t2), the apparent diffusion coefficient image (ADC) and the diffusion image (Diff). All image sequences were obtained and stored using the standard DICOM 3.0. Besides, all studies were acquired using multiple 1.5T scanners and all contrast agents

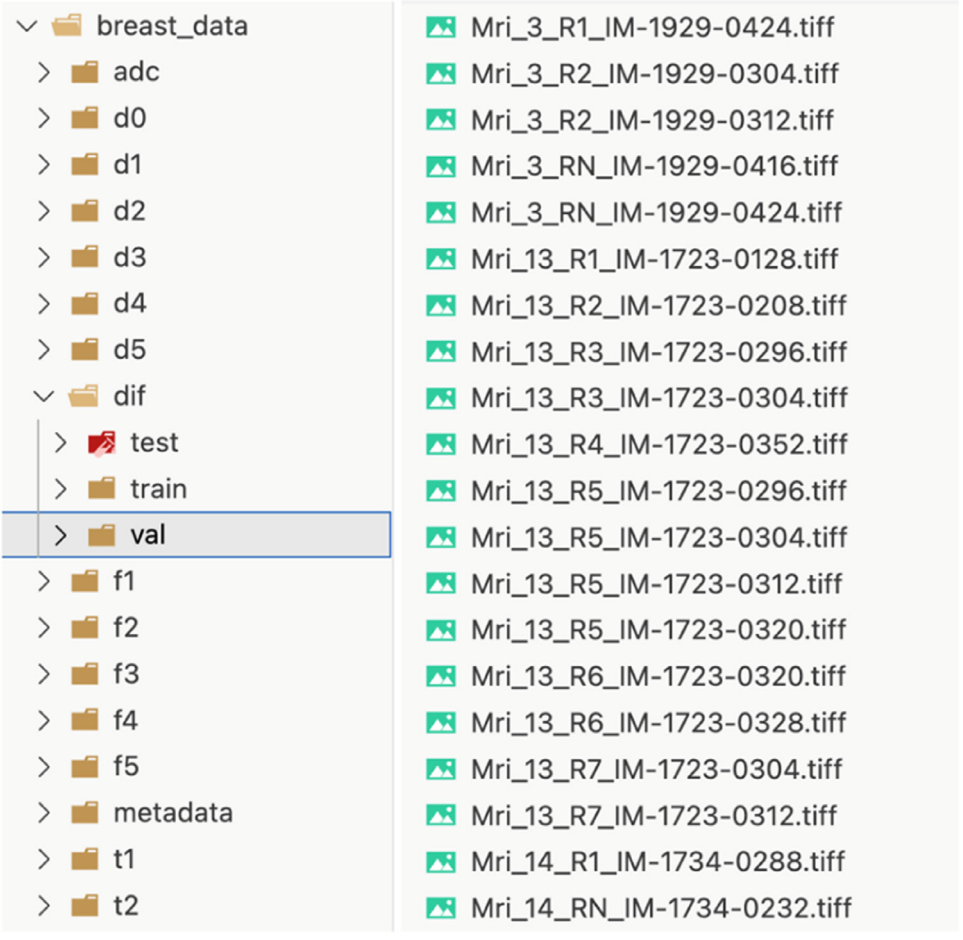

Fig. 1. Tree structure of the dataset. Each folder contains the files corresponding to its image sequence. Besides, each folder contains three more folders regarding the train, test and validation partitions. Image files are stored in tiff format to preserve their original values without compression instead of DICOM files to reduce required storage space.

were gadolinium-based with dosages between 0.014 and 0.016 l/mol as they were obtained retrospectively.

Each study was annotated with multiple visual lesions in the tissue (benign and malignant), and its location was provided using the convention in Table 2. For image postprocessing after acquisition, we computed the phase postcontrast sequences manually using the software Horos[1]. Then, images were stored and treated individually as independent sequences. For image preprocessing, all volume images were filtered and only slices with any ROI were considered. The slice of the annotation is located through the Center_z parameter in the sequences and images were taken separately according to their visualization. Each image was stored and marked in tiff format containing the patient identifier, the ROI identifier and original DCM filename (e.g., Mri_1_R1_IM-1682-0111.tiff). Finally, patients were manually splitted into train, test and validation partitions according to their BI-RADS and American College or Radiology density (ACR). To

[1] Horos is a free and open source code software (FOSS) program that is distributed free of charge under the LGPL license at Horosproject.org and sponsored by Nimble Co LLC d/b/a Purview in Annapolis, MD USA.

Table 2
Description of the metadata provided with the images. The metadata is provided in csv format for each train, test and validation partitions, and each feature is coded in the columns of the files.

Tag	Description
ROI	Region identifier (as there might be multiple ROIs per patient)
Center_x	Center of the ROI in the x axis of coordinates (in pixels).
Center_y	Center of the ROI in the y axis of coordinates (in pixels).
Center_z	Slice where the region is located. Might be useful to locate lesion in the volume.
Distance_x	Distance from the center to the left side of the lesion in the x axis (in pixels).
Distance_y	Distance from the center to the upper side of the lesion in the x axis (in pixels).
ACR	Density of tissue, according to the American College of Radiology (ACR) scale
BI-RADS	BI-RADS category per ROI

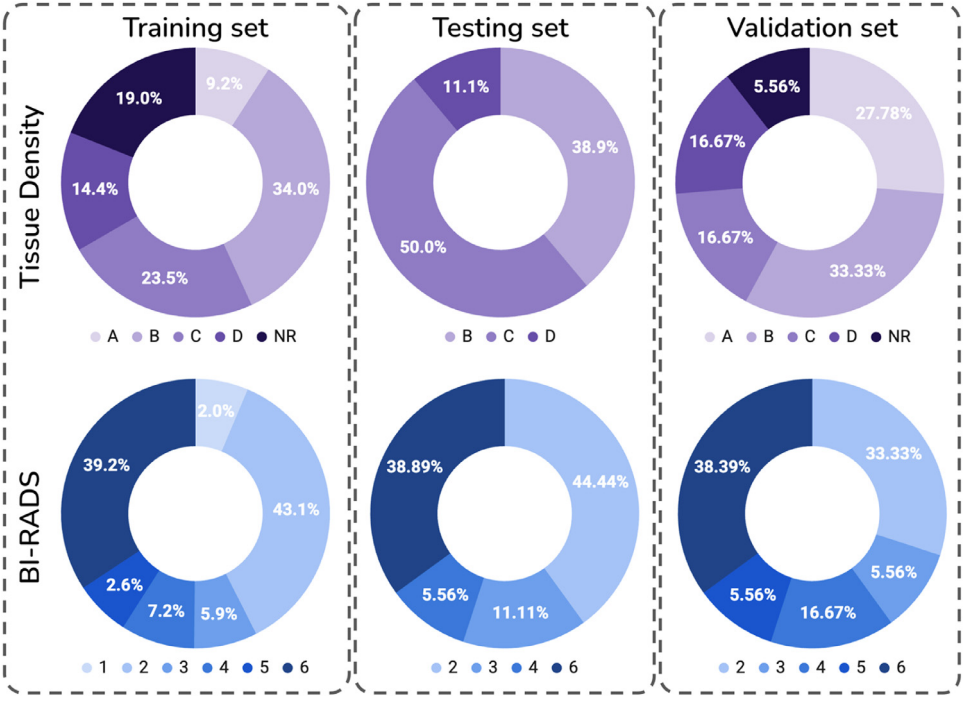

Fig. 2. Distribution of the dataset in terms of the tissue density and their BI-RADS category.

perform the selection, test and validation partitions were prioritized to be balanced in terms of benign and malignant lesions for analysis purposes. A similar procedure was also considered for the ACR parameter, where tissue densities A and B were considered as mostly fatty tissue; and tissue densities C and D were considered as mostly fibroglandular tissue. Fig. 2 shows the final distribution of the three train, test and validation set in terms of the ACR and BI-RADS.

Limitations

Despite the dataset belongs to Latin American patients, no demographic information regarding location (city), age, race, socioeconomical status, previous medical history or other useful data was not found. Besides, no information regarding treatment or outcome of patients was not available neither. Thereupon, other information related to this issue was discarded.

Ethics Statement

Authors of this work confirm that an informed consent form was provided and signed to participants and their data was fully anonymized to preserve their privacy. No additional biological risks to patients or the environment are associated with the acquisition of data or similar.

The authors confirm that they have read and follow the ethical requirements for publication in Data in Brief and confirm that the current work does not involve human subjects, animal experiments, or any data collected from social media platforms.

Data Availability

LA-Breast DCE-MRI Dataset (Original data) (Mendeley Data)

CRediT Author Statement

Rubén D. Fonnegra: Methodology, Software, Data curation, Writing – original draft; **Carlos Mera:** Conceptualization, Methodology, Validation, Writing – original draft; **Gloria M. Díaz:** Methodology, Validation, Writing – review & editing; **Liliana Hernández:** Conceptualization, Validation, Formal analysis.

Acknowledgements

This work was supported by the Instituto Tecnológico Metropolitano, the Institución Universitaria Pascual Bravo, Ayudas diagnósticas SURA and Agencia de Educación Postsecundaria – Sapiencia, research project grant P20213. This subset of data was compiled from the dataset obtained in the project grant RC 740, Minciencias. This work was carried out while the author Carlos Mera was affiliated with the Instituto Tecnológico Metropolitano. However, the final version of this manuscript was written when the author was already affiliated with the Universidad de Antioquia.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] M.A. Marino, D. Avendano, T. Helbich, K. Pinker, Breast MRI: multiparametric and advanced techniques, Breast Imaging Diagn. Intervention (2022) 231–257.
[2] R.D. Fonnegra, M.L. Hernandez, J.C. Caicedo, G.M. Díaz, Early-to-Late prediction of DCE-MRI contrast-enhanced images in using generative adversarial networks, in: 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI), IEEE, 2023, pp. 1–5.
[3] S.C. Partridge, K.M. Stone, R.M. Strigel, W.B. DeMartini, S. Peacock, C.D. Lehman, Breast DCE-MRI: influence of post-contrast timing on automated lesion kinetics assessments and discrimination of benign and malignant lesions, Acad. Radiol. 21 (9) (2014) 1195–1203.
[4] A.M. Hasan, N.K. Al-Waely, H.K. Ajobouri, R.W. Ibrahim, H.A. Jalab, F. Meziane, A classification model of breast masses in DCE-MRI using kinetic curves features with quantum-Raina's polynomial based fusion, Biomed. Signal Process. Control 84 (2023) 105002.
[5] S. Coimbra, S. Rocha, N.R. Sousa, C. Catarino, L. Belo, E. Bronze-da-Rocha, … A. Santos-Silva, Toxicity mechanisms of gadolinium and gadolinium-based contrast agents—A review, Int. J. Mol. Sci. 25 (7) (2024) 4071.