Data in Brief 57 (2024) 110980

Contents lists available at ScienceDirect

Data in Brief

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

Data Article

# A dataset of mammography images with area-based breast density values, breast area, and dense tissue segmentation masks

Hamid Behravan[a,*], Naga Raju Gudhe[a], Hidemi Okuma[b], Mazen Sudah[b], Arto Mannermaa[a,c]

[a] Institute of Clinical Medicine, Pathology and Forensic Medicine, Multidisciplinary Cancer Research Community RC Cancer, University of Eastern Finland, P.O. Box 1627, 70211 Kuopio, Finland
[b] Department of Clinical Radiology, Kuopio University Hospital, P.O. Box 100, Kuopio 70029, Finland
[c] Biobank of Eastern Finland, Kuopio University Hospital, Kuopio, Finland

ARTICLE INFO

*Article history:*
Received 18 April 2024
Revised 30 May 2024
Accepted 24 September 2024
Available online 30 September 2024

Dataset link: Mammogram Density Assessment Dataset (Original data)

*Keywords:*
Breast density estimation
Mammogram assessment
Image segmentation
Dense tissues

ABSTRACT

A new dataset is presented to propel research in automated breast density estimation, a crucial factor in mammogram interpretation. Mammography, a low-dose X-ray technique for breast cancer screening, can be affected by breast density. Dense tissue appears white on mammograms, potentially obscuring tumors. This dataset, built upon the public VinDr-Mammo dataset, offers 745 mammogram images (including training and test sets) along with expert-radiologist annotations for both the entire breast and dense tissue regions. Researchers can leverage this dataset for multiple purposes: training deep learning models for automated breast density analysis, refining segmentation methods for accurate delineation of breast tissue, and benchmarking existing and novel breast density estimation algorithms. This resource holds promise for improving breast cancer screening through advancements in automated breast density analysis.



* Corresponding author.
*E-mail address:* hamid.behravan@uef.fi (H. Behravan).

https://doi.org/10.1016/j.dib.2024.110980

Specifications Table

| | |
|---|---|
| Subject | Health and medical sciences |
| Specific subject area | Medical imaging: Image segmentation and breast density estimation for improved breast cancer screening. |
| Type of data | Analyzed<br>Formats: jpg, csv |
| Data collection | Mammograms were sourced from the public VinDr-Mammo dataset. We randomly selected 745 mammograms and partitioned the data into a training set of 596 mammograms and a separate test set of 149 images. The mammograms' names have been preserved to enable the linkage of segmentation masks and/or mammograms back to their original DICOM images from the VinDr-Mammo dataset.An expert radiologist annotated each mammogram with breast area and dense tissue masks and provided area-based breast density percentage for each mammogram. We used the following steps to create the binary masks:<br>• The breast area is segmented with VGG image annotator. We removed the background by overlapping the masks and the images. We normalized the breast intensity with min-max scaler.<br>• The dense tissues are segmented with an image thresholding technique implemented in Python 3.6 and Flask. |
| Data source location | VinDr-Mammo raw dataset: Smart Health Center – VinBigData JSC, 9th floor, Century Tower, Times City, 458 Minh Khai, Hai Ba Trung, Ha Noi.<br>https://vindr.ai/datasets/mammo |
| Data accessibility | Repository name: Mendeley<br>Data identification number: 10.17632/tdx3h2fn9v.4<br>Direct URL to data: https://data.mendeley.com/datasets/tdx3h2fn9v/4 |
| Related research article | Gudhe, N.R., Behravan, H., Sudah, M. et al. Area-based breast percentage density estimation in mammograms using weight-adaptive multitask learning. Sci Rep 12, 12,060 (2022). Doi: 10.1038/s41598-022-16141-2 [1] |

## 1. Value of the Data

- Breast Density Estimation: This dataset facilitates the development of automated methods for breast density analysis, potentially overcoming limitations of subjective visual assessment.
- Clinical Application: Supports for developing computer-aided diagnosis (CAD) tools for physician-aided breast cancer screening.
- Segmentation Methods: Improve techniques for medical image segmentation through accurate delineation of breast and dense tissue regions.
- Algorithm Benchmarking: Compare existing and new image segmentation and breast density estimation methods.

## 2. Background

This dataset was compiled to address the limitations of current methods for breast density assessment in mammograms, especially the challenges of:

- Shortage of radiologists: There are not enough radiologists to efficiently analyze the large number of mammograms needed for screening.
- Subjectivity: Radiologist assessments of breast density can vary, leading to inconsistencies.
- Limitations of existing tools: Current CAD tools for breast density estimation often have limitations, such as restricted functionality to specific mammogram views and difficulties with accurate segmentation.

This dataset offers a unique solution by expanding the original mammogram images with:

- Binary masks of the breast area: These expert-annotated masks precisely delineate the entire breast region in each mammogram, providing valuable ground truth data for segmentation methods.
- Binary masks of dense tissue: Similarly, these masks accurately identify areas of dense tissue within each mammogram, further enhancing the dataset's utility for training and evaluating segmentation algorithms.

The dataset facilitates the development of automated breast density estimation with deep learning. It also serves as a valuable tool for researchers developing and benchmarking medical image segmentation methods specifically focused on breast tissue analysis in mammograms.

## 3. Data Description

This dataset consists of segmentation masks for dense tissue and breast area as well as area-based breast density percentage values from the VinDr-Mammo public dataset accessible from [4]. All annotations were performed and validated by an expert radiologist.

**Files**:

The data is provided in two compressed archives, '*train.zip*' and '*test.zip*'.

- **train.zip**: This archive contains two sub-folders:
  - **breast_masks**: This sub-folder contains the ground truth segmentation masks for the breast area, also in JPG format.
  - **dense_masks**: This sub-folder contains the ground truth segmentation masks for the dense tissue, again in JPG format.

The segmentation masks have the dimensions of 2800 $\times$ 3518 pixels.

**File Lists**:

Two CSV files are provided alongside the compressed archives:

- **train.csv**: This file contains information about the training set with two columns:
  - **Filename**: This column contains the filenames of the training set images. These images can be directly downloaded from the VinDr-Mammo dataset, https://physionet.org/content/vindr-mammo/1.0.0/.
  - **Density**: This column provides the ground truth continuous breast density value for each mammogram in the training set, intended for the breast density estimation task.
- **test.csv**: This file contains a single column, "*Filename*", listing the filenames of the test set. No ground truth information is provided for the test set. Ground truths are intentionally kept private for Breast Density Kaggle Challenge https://www.kaggle.com/competitions/breast-density-prediction, however, will be eventually open to public in the dataset repository.

Fig. 1 presents examples of breast mammogram images in both craniocaudal (CC) and mediolateral oblique (MLO) views. Each image is accompanied by two corresponding binary masks: one for the breast area and another for dense tissue segmentation. The ground truth breast density value for each mammogram is also provided. Distribution of density values in the training and test sets are shown in Figs. 2 and 3, respectively. Table 1 showcases the distribution of mammogram characteristics across the training and test splits obtained from the VinDr-Mammo dataset. It details the BIRADS categories, view positions (CC and MLO), density categories (A, B, C, D), and laterality (left or right) of the selected mammograms in this study.

## 4. Experimental Design, Materials and Methods

The VinDr-Mammo dataset contains information on 5000 mammograms taken between 2018 and 2020 stored in the Picture Archiving and Communication Systems (PACS) of HMUH and

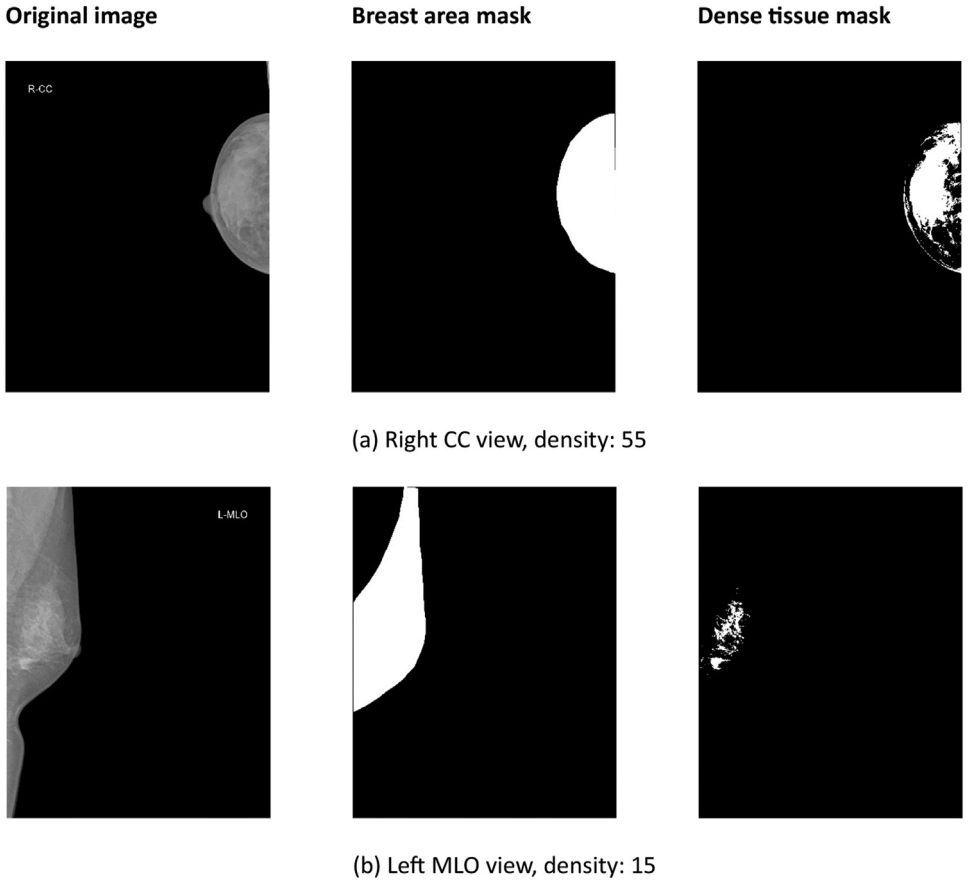


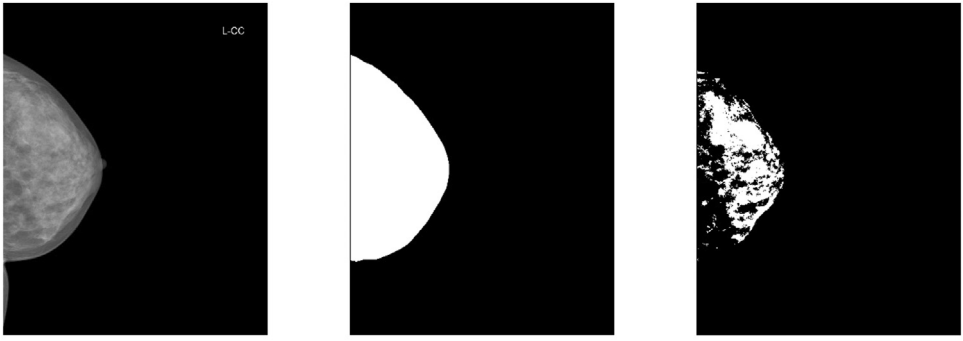

**Fig. 1.** Examples of mammogram images with their corresponding breast area and dense tissue segmentation masks. Each mammogram is assigned a breast density value given by an expert radiologist.

H108 [2,3]. Each mammogram includes four standard views and has been interpreted by two radiologists. Any discrepancies in their interpretations were resolved through arbitration. The data is intended to be used for evaluating the Breast Imaging Reporting and Data System (BI-RADS) and breast density. In terms of overall breast assessment, BI-RADS assessment categories (from 1 to 5) and breast density levels (A, B, C, or D) are provided for each mammogram.

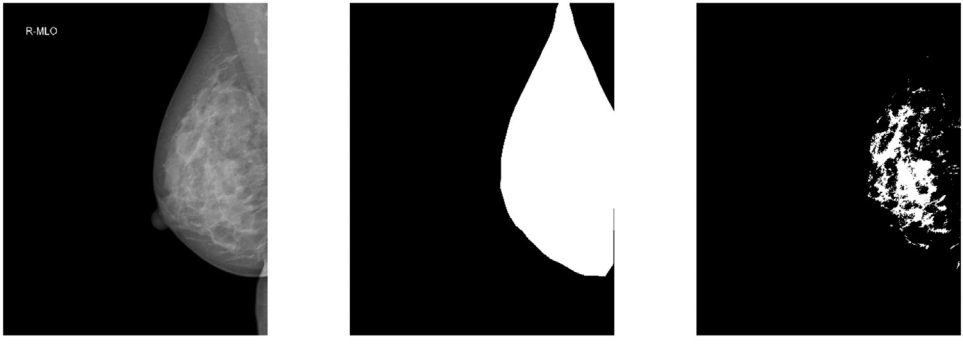
### *4.1. Generating ground-truth annotations*

We utilized the VGG Image Annotation Software [5] to manually annotate the breast area. These annotations were then converted into JSON format. Subsequently, using the OpenCV 4.9.0 python package, we created the breast area binary mask as illustrated in Fig. 4.

We further developed an in-house dense tissue annotation tool using Python 3.8 and the Django 1.9.8 web framework. Fig. 5 visualizes the screenshot of this tool. The process of dense tissue annotation begins with the conversion of the mammogram to a grayscale image. The radiologist then moves a scroll bar corresponding to different threshold values, marking different white pixels as dense tissues as they are separated from the background. At a specific threshold value, when the desired dense tissue is observed by the radiologist, the generated dense tissue




(c) Left CC view, density: 60




(d) Right MLO view, density: 40

**Fig. 1.** Continued

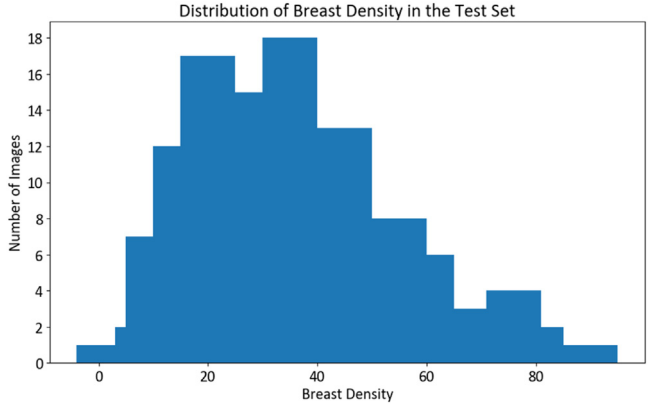



**Fig. 2.** Distribution of breast density values in the training set.

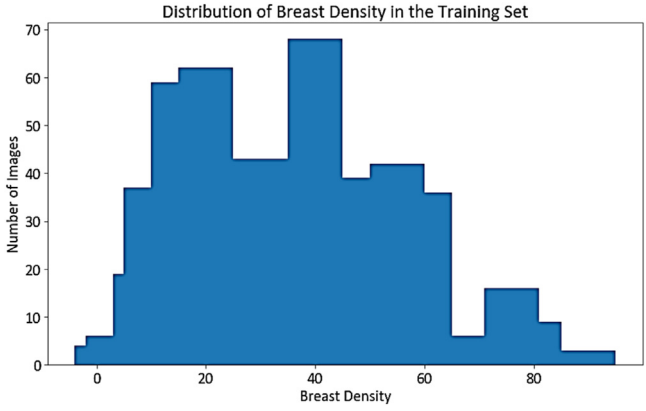



**Fig. 3.** Distribution of breast density values in the test set.

**Table 1**
Distribution of mammogram characteristics across the training and test splits obtained originally from the VinDr-Mammo dataset.

| Split | Laterality | View position | BIRADS | | | | | Density category | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | 1 | 2 | 3 | 4 | 5 | A | B | C | D |
| Train (596) | Left (296) | CC (148) | 107 | 29 | 7 | 4 | 1 | – | 17 | 116 | 15 |
| | | MLO (148) | 108 | 24 | 8 | 7 | 1 | 1 | 16 | 111 | 20 |
| | Right (300) | CC (134) | 83 | 34 | 8 | 6 | 3 | – | 11 | 90 | 33 |
| | | MLO (166) | 117 | 35 | 8 | 5 | 1 | 1 | 14 | 23 | 128 |
| Test (149) | Left (84) | CC (44) | 30 | 9 | 2 | 2 | 1 | – | 3 | 8 | 33 |
| | | MLO (40) | 25 | 12 | 2 | – | 1 | – | | 33 | 7 |
| | Right (65) | CC (29) | 24 | 3 | | 2 | – | – | 1 | 22 | 6 |
| | | MLO (36) | 24 | 8 | 3 | 1 | – | 1 | 3 | 4 | 28 |

binary mask is saved. Subsequently, the generated dense tissue binary mask is overlaid on the original breast mask to remove additional tissues in the dense mask (abdominal, pectoral, see red counters in Fig. 5) and to reduce noise.

Fig. 6 depicts an example of the ground truth annotations of the breast area (marked in red) and dense tissues (highlighted in green), superimposed on the original mammogram image.

We have open-sourced all the code used to convert DICOM images to JPG format, generate the breast area and dense tissue binary masks, as well as the visualization of the segmentations online at https://github.com/uefcancer/Mammography_dataset.

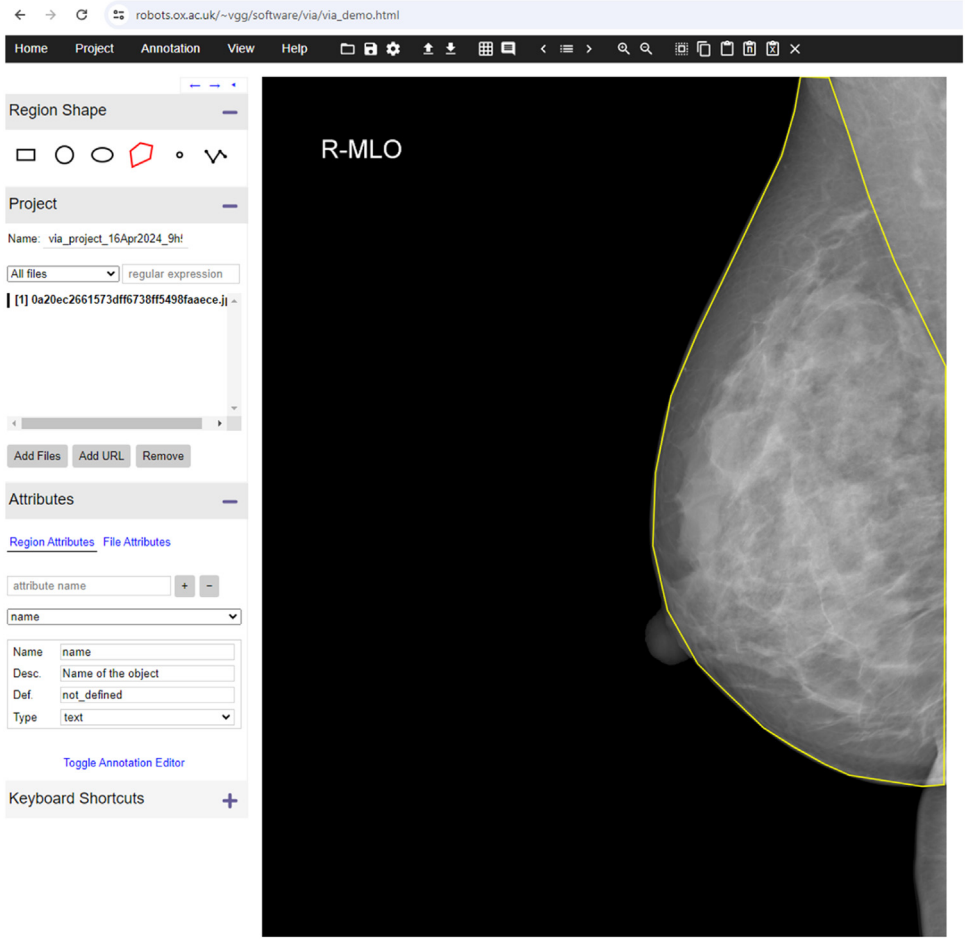



**Fig. 4.** Illustration of the breast area binary mask created using OpenCV 4.9.0, based on manual annotations performed with VGG Image Annotation Software, subsequently converted into JSON format.

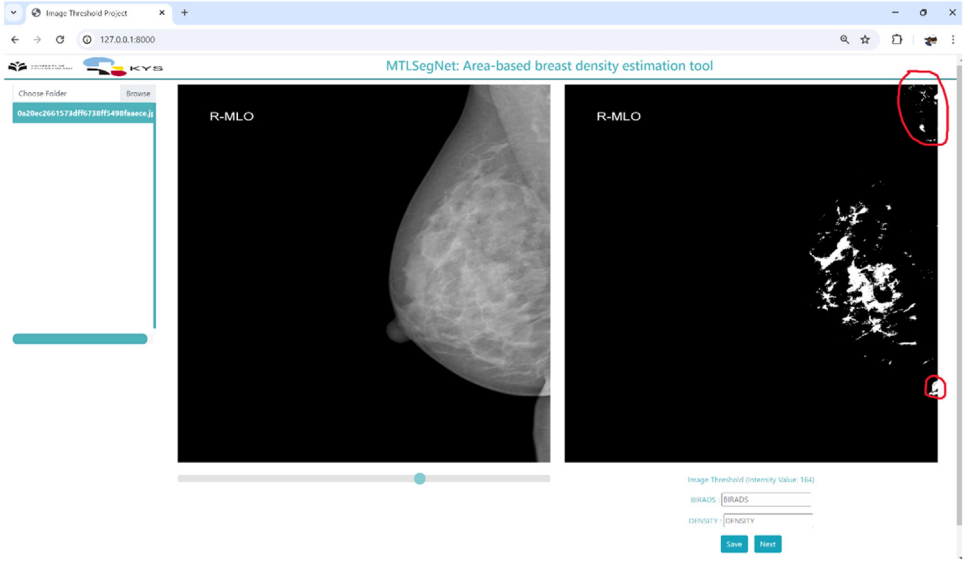



**Fig. 5.** Screenshot of the in-house dense tissue annotation tool developed using Python 3.8 and Django 1.9.8. The figure illustrates the annotation process conducted by an expert radiologist, highlighting the dense tissues within the breast area. The overlaid binary mask of the breast area and the removal of additional noise (depicted as red counters) during the annotation process are also shown. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

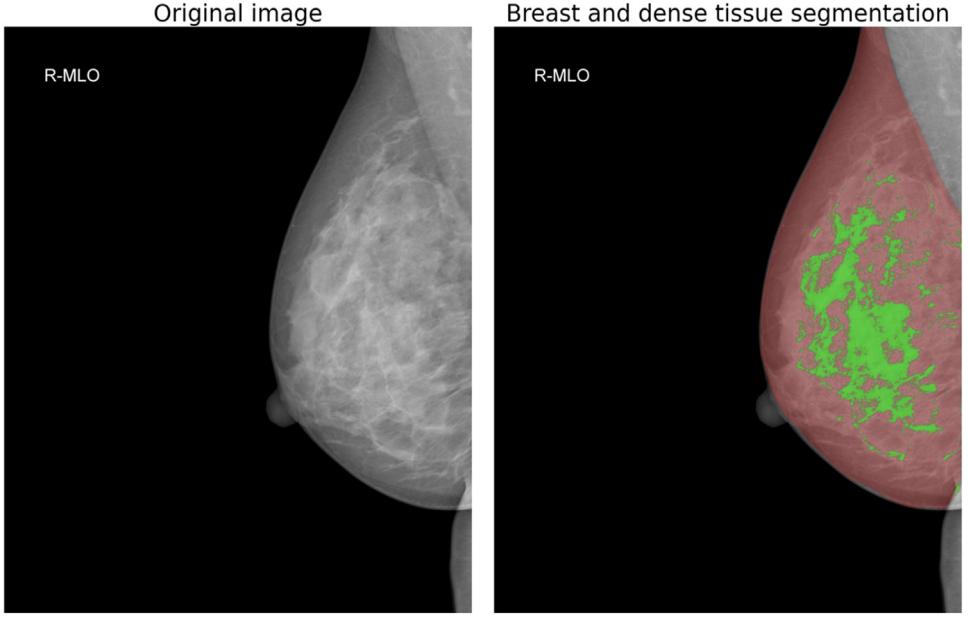



**Fig. 6.** The ground truth annotations of the breast area (marked in red) and dense tissues (highlighted in green), superimposed on the original mammogram image. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

## Limitations

Ground truth binary breast area and dense tissue masks, along with density values for the test set, are not intentionally provided. This data is reserved for evaluation purposes in the breast density challenge hosted on Kaggle https://www.kaggle.com/competitions/breast-density-prediction. However, following the challenge, the test set ground truths will be uploaded to the dataset repository, allowing for a retrospective assessment of model performance on unseen data.

## Ethics Statement

This study didn't conduct experiments involving neither humans nor animals. The VinDr-Mammo dataset was ethically obtained with approval from the Institutional Review Boards of Hanoi Medical University Hospital (HMUH) and Hospital 108 (H108).

## CRediT Author Statement

**Hamid Behravan**: Conceptualization, Methodology, Writing, Original Draft Preparation, Funding. **Naga Raju Gudhe**: Conceptualization, Methodology, Writing, Software. **Hidemi Okuma**: Expert Data Annotation, Validation. **Mazen Sudah**: Expert Data Annotation, Validation. **Arto Mannermaa**: Writing- Reviewing and Editing, Supervision, Funding.

## Data Availability

Mammogram Density Assessment Dataset (Original data) (Mendeley Data).

## Acknowledgments

The authors acknowledge financial support from the Finnish Innovation Fund - Sitra (grant number 29330000451).

## 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] N.R. Gudhe, H. Behravan, M. Sudah, et al., Area-based breast percentage density estimation in mammograms using weight-adaptive multitask learning, Sci. Rep. 12 (2022) 12060, doi:10.1038/s41598-022-16141-2.
[2] H.T. Nguyen, H.Q. Nguyen, H.H. Pham, et al., VinDr-Mammo: a large-scale benchmark dataset for computer-aided diagnosis in full-field digital mammography, Sci. Data 10 (2023) 277, doi:10.1038/s41597-023-02100-7.
[3] A. Goldberger, L. Amaral, L. Glass, J. Hausdorff, P.C. Ivanov, R. Mark, … H.E. Stanley, PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals, Circulation [Online]. 101 (23) (2000) e215–e220.
[4] H. Behravan, N.R. Gudhe, H. Okuma, S. Mazen, A. Mannermaa, Mammogram density assessment dataset, Mendeley Data V5 (2024), doi:10.17632/tdx3h2fn9v.5.
[5] A. Dutta, A. Zisserman, The VIA annotation software for images, audio and video, in: In Proceedings of the 27th ACM International Conference on Multimedia, 2019, pp. 2276–2279, doi:10.1145/3343031.3350535.