Data in Brief 31 (2020) 105928

Contents lists available at ScienceDirect

Data in Brief

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

Data Article

Dataset of breast mammography images with masses

Mei-Ling Huang[a,*], Ting-Yu Lin[a]

[a] Department of Industrial Engineering & Management, National Chin-Yi University of Technology, Taichung, Taiwan

ARTICLE INFO

Article history:
Received 6 May 2020
Revised 29 May 2020
Accepted 22 June 2020
Available online 25 June 2020

Keywords:
Breast mammography images
Data augmentation
Contrast limited adaptive histogram equalization
Breast mass
Breast density

ABSTRACT

Among many cancers, breast cancer is the second most common cause of death in women. Early detection and early treatment reduce breast cancer mortality. Mammography plays an important role in breast cancer screening because it can detect early breast masses or calcification region. One of the drawbacks in breast mammography is breast cancer masses are more difficult to be found in extremely dense breast tissue. We select 106 breast mammography images with masses from INbreast database. Through data augmentation, the number of breast mammography images was increased to 7632. We utilize data augmentation on breast mammography images, and then apply the Convolutional Neural Networks (CNN) models including AlexNet, DenseNet, and ShuffleNet to classify these breast mammography images.

* Corresponding author.
E-mail address: huangml@ncut.edu.tw (M.-L. Huang).

https://doi.org/10.1016/j.dib.2020.105928

Specifications Table

Subject	Medicine and Dentistry
Specific subject area	Radiology and imaging
Type of data	Raw and analyzed
How data were acquired	The data was obtained from Breast center in CHSJ, Porto.
Data format	PNG
Parameters for data collection	Among 410 mammograms in INbreast database, 106 images were breast mass and were selected in this study.
Description of data collection	Through data augmentation, the number of breast mammography images was increased to 7632 in this study.
Data source location	Centro Hospitalar de S. Joao [CHSJ], Breast center, Porto
Data accessibility	http://dx.doi.org/10.17632/x7bvzv6cvr.1

Value of the Data

- Breast density affects the diagnosis of breast cancer. The dataset combines four breast densities with benign or malignant status to become eight groups for breast mammography images.
- The dataset helps physicians for early detection and treatment to reduce breast cancer mortality.
- The numbers of images in the dataset are increased through data augmentation. It allows the model to learn more pictures of different situations and angles to accurately classify new images.
- Different machine learning and deep learning algorithms can be used to model the data and predict the classification results.

1. Data description

Mammography images of INbreast database was originally collected from Centro Hospitalar de S. Joao [CHSJ], Breast center, Porto. INbreast database collects data from Aug. 2008 to July 2010, which contains 115 cases with a total of 410 images [1]. Among them, 90 cases were women with disease on both breasts. There are four different types of breast diseases recorded in the database, including Mass, Calcification, Asymmetries, and Distortions. The images of this database have two perspectives of Craniocaudal (CC) and medilateral oblique (MLO), and the breast density is divided into four categories according to BI-RADS standards [2], which are Entirely fat (Density 1), Scattered fibroglandular densities (Density 2), Heterogeneously dense (Density 3), and Extremely dense (Density 4). Imagegs were saved in two sizes: 3328 X 4084 or 2560 X 3328 pixels in DICOM [2].

Among 410 mammograms in INbreast database, 106 images were breast mass and were selected in this study. Through data augmentation, the number of breast mammography images was increased to 7632 in this study. Fig. 1 presents examples of breast mammography images with masses for four density categories with benign or malignant status: (a) Density 1 with breast mass Benign; (b) Density 1 with breast mass Malignant; (c) Density 2 with breast mass Benign; (d) Density 2 with breast mass Malignant; (e) Density 3 with breast mass Benign; (f) Density 3 with breast mass Malignant; (g) Density 4 with breast mass Benign; (h) Density 4 with breast mass Malignant. Compared to benign masses, the shapes of malignant masses are irregular.

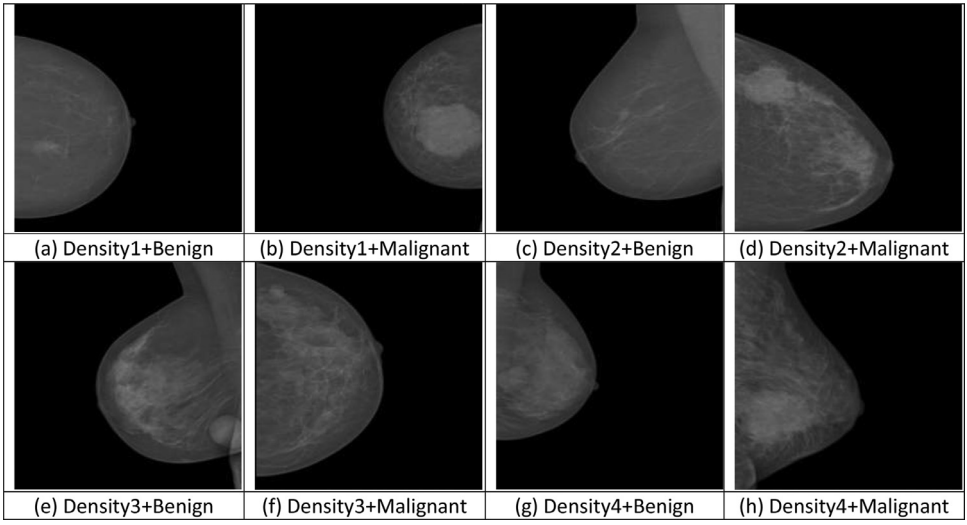

Fig. 1. Breast Masses for four density categories with benign or malignant status.

Table 1
Number of images for breast density with benign and malignant class labels.

	Category	Number
(a)	Density1+Benign	12
(b)	Density1+Malignant	30
(c)	Density2+Benign	4
(d)	Density2+Malignant	32
(e)	Density3+Benign	13
(f)	Density3+Malignant	8
(g)	Density4+Benign	6
(h)	Density4+Malignant	1
	Total	106

2. Experimental design, materials, and methods

2.1. Data collection

Each image was marked with its corresponding breast density and the original images in INbreast database are DICOM files. We converted the DICOM files to PNG files through Matlab R2019a [3].

Combining four breast density categories and breast benign or malignant status, therefore, there are 8 categories in our classification task. The eight categories are:

(1) The category of breast density is 1 and breast mass is benign (Density1+Benign)
(2) The category of breast density is 1 and breast mass is malignant (Density1+Malignant)
(3) The category of breast density is 2 and breast mass is benign (Density2+Benign)
(4) The category of breast density is 2 and breast mass is malignant (Density2+Malignant)
(5) The category of breast density is 3 and breast mass is benign (Density3+Benign)
(6) The category of breast density is 3 and breast mass is malignant (Density3+Malignant)
(7) The category of breast density is 4 and breast mass is benign (Density4+Benign)
(8) The category of breast density is 4 and breast mass is malignant (Density4+Malignant).

Table 1 displays the number of images selected from INbreast dataset for each breast density with bening or malignant class labels. The number of (a) Density 1 with breast mass Benign; (b)

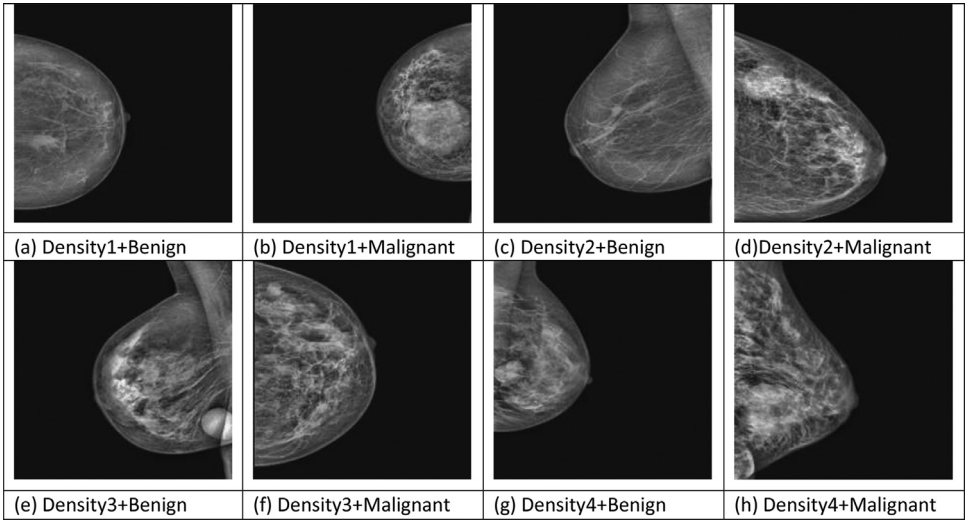

Fig. 2. The image after CLAHE processing.

Table 2
Number of images before image augmentation.

Category		Image Before Data Augmentation		
		All	Training	Testing
1	Density1+Benign	24	19	5
2	Density1+Malignant	60	48	12
3	Density2+Benign	8	6	2
4	Density2+Malignant	64	51	13
5	Density3+Benign	26	2	5
6	Density3+Malignant	16	13	3
7	Density4+Benign	12	10	2
8	Density4+Malignant	2	2	0
Total		212	170	42

Density 1 with breast mass Malignant; (c) Density 2 with breast mass Benign; (d) Density 2 with breast mass Malignant; (e) Density 3 with breast mass Benign; (f) Density 3 with breast mass Malignant; (g) Density 4 with breast mass Benign; (h) Density 4 with breast mass Malignant are 12, 30, 4, 32, 13, 8, 6, and 1, respectively.

2.2. Pre-processing

The image preprocessing method contrast limited adaptive histogram equalization (CLAHE) was used on the original 106 images. Fig. 1 presents examples of breast mammography images with masses for four categories after CLAHE processing: (a) Density 1 with breast mass Benign; (b) Density 1 with breast mass Malignant; (c) Density 2 with breast mass Benign; (d) Density 2 with breast mass Malignant; (e) Density 3 with breast mass Benign; (f) Density 3 with breast mass Malignant; (g) Density 4 with breast mass Benign; (h) Density 4 with breast mass Malignant. Compared with Fig. 1, it can be seen from Fig. 2 that the mass location of the image after CLAHE processing is clearer than the original image. We have 106 original images and another 106 images after CLAHE processing, so there are $106 * 2 = 212$ images. Table 2 presents the number of images for 8 categories in training and testing sets after CLAHE processing.

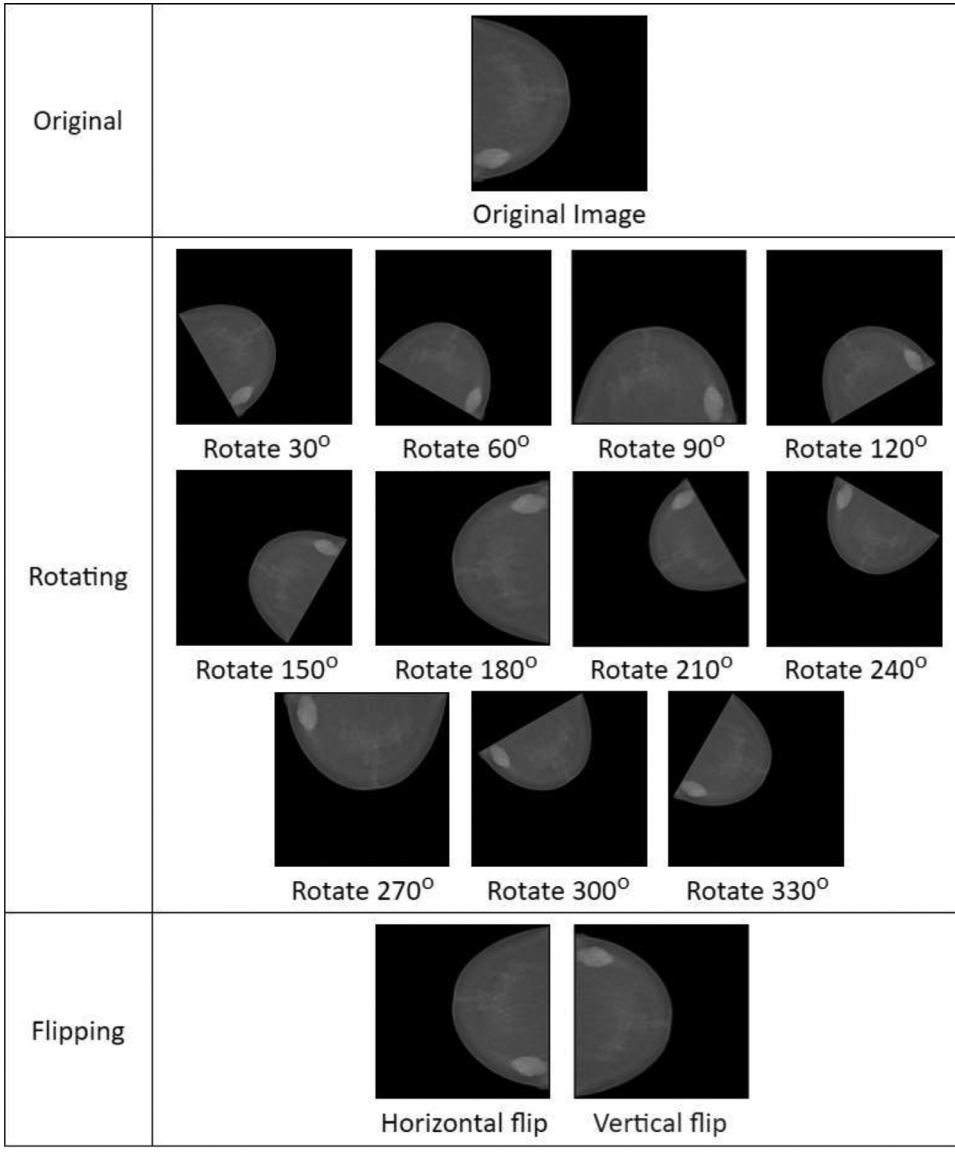

Fig. 3. Example of original image and images after data augmentation.

2.3. Data augmentation

In addition to CLAHE, we further perform data augmentation with multi-angle rotation ($\theta = 30, 60, 90, 120, 150, 180, 210, 240, 270, 300, 330^{\circ}$), and then flips the original image and 11 angle rotation images horizontally and vertically. The method not only increases the number of samples, but also prevents the problem of overfitting. Fig. 3 is an example to show the original image, images with multi-angle rotation, and images with horizontally and vertically flipping.

The number of images after image augmentation is 7632. The number of images in the training set and testing set for 8 categories after image augmentation are shown in Table 3.

Table 3
Number of images after image augmentation.

Category		Image After Data Augmentation		
		All	Training	Testing
1	Density1+Benign	864	691	173
2	Density1+Malignant	2160	1728	432
3	Density2+Benign	288	230	58
4	Density2+Malignant	2304	1843	461
5	Density3+Benign	936	749	187
6	Density3+Malignant	576	461	115
7	Density4+Benign	432	346	86
8	Density4+Malignant	72	58	14
Total		7632	6106	1526

The dataset in this study was built to be used in convolutional neural network including AlexNet, DenseNet, and ShuffleNet for the classification of benign and malignant mammograms. Due to different image sizes required by different CNN models, we resized the original images from 3328 x 4084 and 2560 x 3328 pixels into 224×224 pixels for ShuffleNet and DenseNet, and 227×227 pixels for AlexNet.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships which have, or could be perceived to have, influenced the work reported in this article.

Acknowledgments

The authors gratefully acknowledge the financial support of the Ministry of Science and Technology of Taiwan, R.O.C. through its grants MOST 108–2221-E-167–001.

Ethics Statement

This study didn't conduct experiments involving neither humans nor animals.

References

[1] "INbreast Database." http://medicalresearch.inescporto.pt/breastresearch/index.php/Get_INbreast_Database
[2] I.C. Moreira, I. Amaral, I. Domingues, A. Cardoso, M.J. Cardoso, J.S. Cardoso, INbreast: toward a full-field digital mammographic database, Acad Radiol 19 (2) (Feb 2012) 236–248.
[3] *The MathWorks,* Inc. Matlab R2019a. Available: https://www.mathworks.com/products/new_products/release2019a.html