Breast Density Classification with Deep Convolutional Neural Networks

9 authors, including:

Krzysztof Geras
New York University
55 PUBLICATIONS **968** CITATIONS

Yiqiu Shen
New York University
21 PUBLICATIONS **460** CITATIONS

Eric Kim
indipedent
184 PUBLICATIONS **2,608** CITATIONS

Linda Moy
NYU Langone Medical Center
257 PUBLICATIONS **6,209** CITATIONS

Some of the authors of this publication are also working on these related projects:

Project — Diffuson MRI Assessment of Breast Cancer View project

Project — Deep Learning for Breast Cancer Diagnosis View project

BREAST DENSITY CLASSIFICATION WITH DEEP CONVOLUTIONAL NEURAL NETWORKS

Nan Wu[1], Krzysztof J. Geras[1], Yiqiu Shen[1], Jingyi Su[1], S. Gene Kim[3,4], Eric Kim[3], Stacey Wolfson[3], Linda Moy[3,4] & Kyunghyun Cho[1,2]

[1]Center for Data Science, New York University
[2]Courant Institute of Mathematical Sciences, New York University
[3]Center for Biomedical Imaging, Radiology, NYU School of Medicine
[4]Perlmutter Cancer Center, NYU Langone Medical Center

ABSTRACT

Breast density classification is an essential part of breast cancer screening. Although a lot of prior work considered this problem as a task for learning algorithms, to our knowledge, all of them used small and not clinically realistic data both for training and evaluation of their models. In this work, we explore the limits of this task with a data set coming from over 200,000 breast cancer screening exams. We use this data to train and evaluate a strong convolutional neural network classifier. In a reader study, we find that our model can perform this task comparably to a human expert.

Index Terms— convolutional neural networks, deep learning, mammography, breast cancer screening, breast density

1. INTRODUCTION

Although convolutional neural networks (CNNs) are highly successful in a variety of applications [1], they have received relatively little attention in medical image analysis until recently. This has been primarily due to the lack of availability of large public data sets. One of the significant areas for development in medical image analysis is breast cancer screening. Even though performing full diagnosis by the means of a neural network remains a challenge [2], elements of achieving this wider goal are feasible with current state of the art methods. In this paper, we explore one of them, namely breast density classification.

Mammographic density reflects the composition of fibroglandular and fat tissue of a breast as seen on a mammogram. In clinical practice in the United States, breast density is qualitatively categorized into four types: a) almost entirely fatty, b) scattered areas of fibroglandular density, c) heterogeneously dense and d) extremely dense. The last two categories are considered "dense" [3]. Dense breast tissue is common and is the typical fibroglandular breast tissue pattern in young women. About 50% of women over the age of 40

years have dense breasts. After menopause the breasts tend to contain more fat as the glands involute. Dense breast tissue reduces the effectiveness of mammography because it has a "masking effect" and will hide an underlying tumor. Studies also consistently show an increased risk of developing breast cancer in women with high mammographic density compared with women with low mammographic density.

Masking of cancer by dense tissue has become a political issue with women requesting supplemental tests if they have dense breasts. There are currently states in the US that mandate women receive notification about breast density with their mammography results. Due to significant variability in the radiologist' assessment of breast density, computer methods have been developed to improve consistency. One of them, Cumulus, is a software program requiring manual input to outline and measure the area of breast tissue relative to overall breast area [4, 5]. Recently, automated programs have been developed that measure percent density as a function of area or volume. Several automated density programs have demonstrated high reproducibility [6] and correlation with volumetric density as measured by MRI [7]. However, these commercially available products do not involve any learning, therefore they lack the flexibility and robustness of learning models. On the other hand, all learning models in literature were trained and tested with small amounts of data. Hence there still is a need to have a precise automated assessment of breast density based on learning directly from data.

2. DATA

We used a clinically realistic data set of over 200,000 screening mammography exams, each containing at least four images corresponding to the standard four views used in screening mammography [2]. For the purposes of this research, we supplemented this data with labels corresponding to breast density, which we automatically extracted from the textual reports associated with the exams in our data set. A small number of exams in our original data did not have the in-

formation on breast density in the corresponding textual report. We excluded 519 such exams from the data set, which left us with the total of 201,179 exams, containing 19,939 class 0, 85,665 class 1, 83,852 class 2 and 11,723 class 3 exams. Interestingly, analysis of our data confirms that women who were assigned overall BI-RADS 0 ("incomplete") label in their screening mammography tend to have denser breasts than the ones who were assigned BI-RADS 1 ("normal") and BI-RADS 2 ("benign") labels (cf. Table 1).
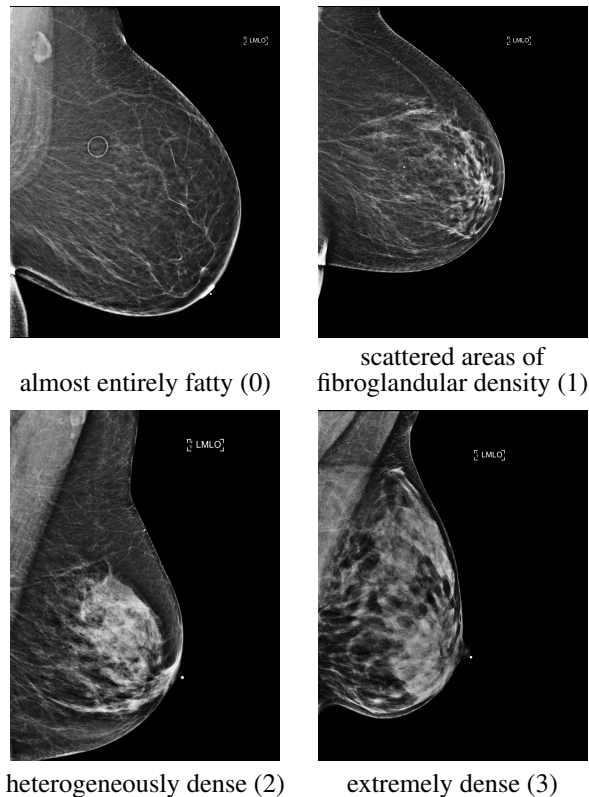

almost entirely fatty (0)

scattered areas of fibroglandular density (1)

heterogeneously dense (2)

extremely dense (3)

Fig. 1. Examples of the four breast density classes.

Table 1. Distribution of labels in our data set. The numbers in the bottom row are numbers of exams falling into different breast density categories. The numbers in the rightmost column are the numbers of exams falling into different overall BI-RADS classes.

		breast density category				
		0	1	2	3	
BI-RADS	0	1702	9607	12656	1839	25804
	1	9803	40060	37167	5157	92187
	2	8434	35998	34029	4727	83188
		19939	85665	83852	11723	

3. MODELS

3.1. Baselines

The most common method to perform the task of breast density prediction in literature is training a classifier with features based on histograms of pixel intensity in the image [8]. This simple method is surprisingly effective. The reason why it works can be easily understood. The difference between pixel intensity occurs because mammograms with a predominance of fat appear darker than the ones with a fibroglandular tissue. This is because this type of tissue absorbs much of the radiation whereas the adipose tissue allows the radiation to get through more easily.

In this work, we used such a model as a baseline. We took pixel intensity histograms as features and used softmax regression as a classifier. For each of the four standard views (L-CC, R-CC, L-MLO and R-MLO) separately, we split intensity values into a set of predefined bins, we normalized them such that frequencies of different bins sum to one and then concatenate such feature vectors. Additionally, to make this model more flexible, we also trained a version of it with an extra hidden layer with 100 hidden units between the input and the softmax regression layer. The hidden layer used rectifier linear function as an activation function.

3.2. Deep convolutional neural network

We used a multi-column deep convolutional neural network of an architecture loosely inspired by the earlier work of Simonyan et al. [9]. The input to the network is four 2600×2000 images corresponding to the standard views used in screening mammography. It is very similar to the architecture in [2] with the exception of the number of the outputs in the softmax layer, since the breast density classification is a four-way classification problem (and not a three-way classification problem as the overall BI-RADS classification). We keep the same values of the hyperparameters too. An overview of our architecture is shown in Figure 2.
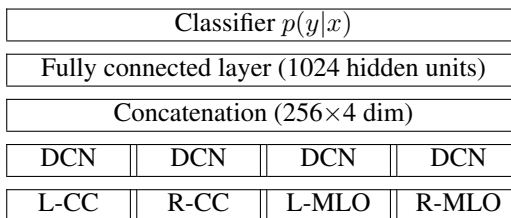
Classifier $p(y|x)$			
Fully connected layer (1024 hidden units)			
Concatenation (256×4 dim)			
DCN	DCN	DCN	DCN
L-CC	R-CC	L-MLO	R-MLO

Fig. 2. An overview of the structure of the convolutional neural network used in our experiments. DCN in the figure above denotes a series of convolutional and pooling layers. L-CC, R-CC, L-MLO and R-MLO refer to inputs corresponding to the four standard views in screening mammography.

4. EXPERIMENTS

We sort the patients according to the date of their latest exam and divide them into training (first 80%), validation (next 10%) and test (last 10%) sets. For the test phase, we only keep the most recent exam for each patient. This way of partitioning the data allows us to estimate performance of our classifiers on future data accurately. During experimentation, we follow the experimental protocol in [2], otherwise we state it explicitly when we deviate from it. In all experiments, we use data augmentation only during training. During validation and test phases all augmentations are off. For all models, we picked the best epoch according to the accuracy on the validation set.

In the baseline model based on histogram features, we use the Adam algorithm with the initial learning rate of 10^{-3}. We tuned the number of bins of pixel intensity histogram, using the validation data to select between 10, 20, 50 and 100.

4.1. Evaluation metrics

Our primary metric in this work is the standard classification accuracy. As the levels of breast density correspond to relative increases in the amount of fibroglandular tissue, two consecutive labels can be confused even by an experienced radiologist. This is why we also considered *top-k* accuracy. In this metric we consider a prediction is correct if the ground truth is among the k most likely labels predicted. Top-k error is currently a popular performance measure on large scale image classification benchmarks such as ImageNet and Places [10]. Additionally, we also considered accuracy only between the two superclasses: "dense" (classes 2 and 3) versus "not dense" (classes 0 and 1).

Secondly, we evaluate our models with respect to the area under the ROC curve (AUC), which is widely used for measuring the predictive accuracy of binary classification models. This metric indicates the relation between the true positive rate and false positive rate when varying the classification threshold. As AUC can only be computed for binary classification, we compute AUCs for all four binary problems of distinguishing between one of the density categories and the rest of the density categories, and then take the macro average, abbreviated as macAUC.

4.2. Impact of the size of the data set

To explore the effect of data set scale, we trained separate networks on training sets of different sizes; 100%, 10% and 1% of the original training set. The results are shown in Table 2. Interestingly, even though training with more data increases performance in all metrics, the difference is not large. In Figure 3 we show the ROC curves for the model trained with 100% of the data.

Table 2. Performance of our CNNs. The * symbol in the leftmost column indicates that a model was initialized using weights of a previously trained overall BI-RADS classifier (cf. subsection 4.3).

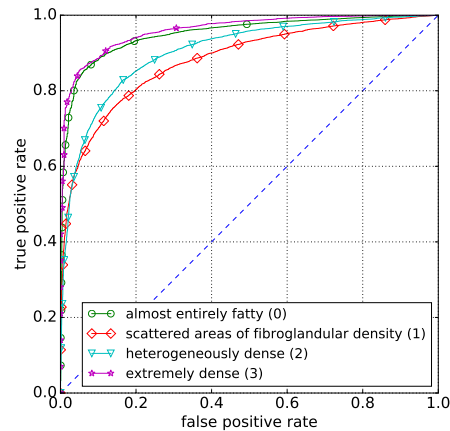| data	macAUC	top-1	top-2	top-3	superclass
1%	0.888	0.729	0.967	0.998	0.849
10%	0.907	0.745	0.976	0.999	0.856
100%	0.916	0.767	0.982	0.999	0.865
*1%	0.892	0.733	0.974	0.998	0.848
*10%	0.909	0.753	0.980	0.998	0.856

Fig. 3. ROC curves for all four classes. The classes 1 and 2 are the hardest for a neural network to distinguish from the rest. The AUC values are 0.955, 0.888, 0.907, 0.960 for classes 0, 1, 2, 3 respectively.

4.3. Transferring knowledge from BIRADS classifier

Transfer learning aims to transfer knowledge between related source and target domains [11]. In computer vision, examples of transfer learning include [12, 13, 14]. The main idea of this technique is to overcome the deficit of training samples by adapting strong classifiers trained for another, related but not identical, task. Considering the amount of parameters in the CNN and the correlation between breast density and overall BI-RADS (cf. Table 1), we applied the idea of transfer learning to accelerate learning of our breast density prediction network. To achieve that, we use the weights of our model previously trained for breast cancer screening [2] to initialize the parameters of the network trained for breast density prediction. The two networks have an identical architecture, with the exception of the softmax layer. This layer of the network trained for breast density prediction is initialized randomly using the recipe from [15].

The models trained with such initialization perform better than their counterparts, trained from scratch in almost all metrics (cf. Table 2), however, only by a small margin. Intriguingly, models initialized with parameters of a previously trained overall BI-RADS classifier achieve the best perfor-

mance in much fewer numbers of training epochs: 20 instead of 50 when using 1% of the original training data 15 instead of 25 when using 10% of the original training data.

4.4. Baseline results

Finally, we trained the baselines based on histograms of pixel intensity. Both versions of the baseline model are trained with 10% of the original training set. The best baseline model without the hidden layer is the one with 20 bins. It achieved 0.832 of macAUC, 67.9% of top-1 accuracy, 90.9% of top-2 accuracy, 99.4% of top-3 accuracy and 81.1% in distinguishing between the "dense" and "not dense" superclasses. The model using 10 bins is the best one for the version with one hidden layer. It achieved 0.842 of macAUC, 69.4% of top-1 accuracy, 90.8% of top-2 accuracy, 99.2% of top-3 accuracy and 82.5% accuracy for superclass classification.

5. COMPARISON TO HUMAN PERFORMANCE

To understand what the limit of performance possible to achieve on this task is, we conducted a reader study with human experts with different levels of experience. The three participants in our reader study were: a medical student (S), a radiology resident (R) and an attending radiologist (A). The experts were all shown the same 100 exams randomly drawn from the test set, each with at least four images corresponding to the standard views used in screening mammography. For each exam, the experts were asked to rank the breast density classes from the most likely to the least likely according to their judgement. The results of this experiment are shown in Table 3. Additionally, we computed analogous values with only two classes instead of the original four: dense breasts (original classes 2 and 3) and not dense breasts (original classes 0 and 1). The results of this experiment are shown in Table 4.

Both human experts and learning models achieve a fair agreement with the labels in the data. Note that the agreement between the predictions of our model and the labels in the data are of similar magnitude to the agreement between the humans themselves.

We also compared our best CNN model to an average of the predictions of human experts. We did that by treating predictions of experts as one-hot vectors and averaging them. In this experiment the humans achieved macAUC of 0.892 (class 0: 0.960, class 1: 0.812, class 2: 0.807 and class 3: 0.990), while the CNN achieved macAUC of 0.934 (class 0: 0.971, class 1: 0.859, class 2: 0.905 and class 3: 1.000).

6. RELATED WORK

Many previous approaches (cf. Table 2 in [16] for a comprehensive review) for this task involve two separate steps of feature extraction and classification. Carneiro et al. [17]

Table 3. Agreement (Cohen's kappa) in choosing the most likely class between different readers (S, R, A), our neural network (N), our baseline (H) and labels in the data set (L).

	L	N	H	S	R	A
L		0.61	0.39	0.41	0.55	0.39
N			0.58	0.53	0.60	0.48
H				0.28	0.37	0.34
S					0.65	0.48
R						0.43

Table 4. Agreement (Cohen's kappa) in distinguishing between dense breasts (classes 2 and 3) and not dense (classes 0 and 1) between different readers (S, R, A), our neural network (N), our baseline (H) and labels in the data set (L).

	L	N	H	S	R	A
L		0.65	0.50	0.50	0.73	0.46
N			0.72	0.62	0.83	0.57
H				0.48	0.69	0.48
S					0.69	0.64
R						0.60

use histograms and Haralick texture descriptors as an input to a multilayer perceptron. In a similar vein of work, Kumar et al. [16] develop an ensemble of six multilayer perceptron networks trained with GLCM mean features. The approaches by Thomaz et al [18] and Fonseca et al [19] apply convolutional neural network too, however, only to extract features for another classification model (MLP and SVM respectively). Additionally, their results are obtained with a much smaller dataset (less than one thousand exams) of low resolution images. To the best of our knowledge, our work is the first end-to-end four-class breast density classifier using supervised deep convolutional neural network on multi-view mammography images. Importantly, we trained and evaluated our models on a diverse clinically realistic data set of high resolution images, approximately two orders of magnitude larger than any previous work we are aware of.

7. CONCLUSIONS

In this work, we trained a deep convolutional neural network classifier using a data set of unprecedented size for the task of breast density classification. The level of agreement between the trained classifier and the classes in the data was found to be similar to that between the human experts and the classes in the data, as well as among the human experts themselves. This result strongly suggests that the proposed classifier may have significant clinical relevance, as it provides quantitative, reproducible prediction, while there is often poor intra-reader and inter-reader correlation in the qualitative assessment of breast density tissue, as was observed in our reader study.

8. REFERENCES

[1] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[2] Krzysztof J. Geras, Stacey Wolfson, Yiqiu Shen, S. Gene Kim, Linda Moy, and Kyunghyun Cho, "High-resolution breast cancer screening with multi-view deep convolutional neural networks," *arXiv:1703.07047v2*, 2017.

[3] Edward A. Sickles, Carl J. D'Orsi, Lawrence W. Bassett, Catherine M. Appleton, Wendie A. Berg, and Elizabeth S. Burnside, "ACR BI-RADS Atlas, Breast Imaging Reporting and Data System," 2013.

[4] Jeffrey W. Byng, Martin J. Yaffe, Roberta A. Jong, Rene S. Shumak, Gina A. Lockwood, David L. Tritchler, and Norman F. Boyd, "Analysis of mammographic density and breast cancer risk from digitized mammograms.," *Radiographics*, vol. 18, no. 6, pp. 1587–1598, 1998.

[5] Wendie A. Berg and Jennifer A. Harvey, "Breast density and supplemental screening," Accessed on 10/27/2017.

[6] Olivier Alonzo-Proulx, Gordon E. Mawdsley, James T. Patrie, Martin J. Yaffe, and Jennifer A. Harvey, "Reliability of automated breast density measurements," *Radiology*, vol. 275, no. 2, pp. 366–376, 2015.

[7] Saskia van Engeland, Peter R. Snoeren, Henkjan Huisman, Carla Boetes, and Nico Karssemeijer, "Volumetric breast density estimation from full-field digital mammograms," *IEEE Transactions on Medical Imaging*, vol. 25, no. 3, pp. 273–282, 2006.

[8] Nico Karssemeijer, "Automated classification of parenchymal patterns in mammograms," *Physics in Medicine & Biology*, vol. 43, no. 2, pp. 365–378, 1998.

[9] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *International Conference on Learning Representations*, 2015.

[10] Maksim Lapin, Matthias Hein, and Bernt Schiele, "Analysis and optimization of loss functions for multiclass, top-k, and multilabel classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[11] Sinno J. Pan and Qiang Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[12] Yusuf Aytar and Andrew Zisserman, "Tabula rasa: model transfer for object category detection," *International Conference on Computer Vision*, 2011.

[13] Tatiana Tommasi, Francesco Orabona, and Barbara Caputo, "Safety in numbers: learning categories from few examples with multi model knowledge transfer," *Conference on Computer Vision and Pattern Recognition*, 2010.

[14] Maxime Oquab, Leon Bottou, Ivan Laptev, and Josef Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," *Conference on Computer Vision and Pattern Recognition*, 2014.

[15] Xavier Glorot and Yoshua Bengio, "Understanding the difficulty of training deep feedforward neural networks," *International Conference on Artificial Intelligence and Statistics*, 2010.

[16] Indrajeet Kumar, H.S. Bhadauria, Jitendra Virmani, and Shruti Thakur, "A classification framework for prediction of breast density using an ensemble of neural network classifiers," *Biocybernetics and Biomedical Engineering*, vol. 37, no. 1, pp. 217–228, 2017.

[17] Pedro C. Carneiro, Marcelo L. N. Franco, Ricardo L. Thomaz, and Ana C. Patrocinio, "Breast density pattern characterization by histogram features and texture descriptors," *Research on Biomedical Engineering*, vol. 33, pp. 69 – 77, 2017.

[18] Ricardo L. Thomaz, Pedro C. Carneiro, and Ana C. Patrocinio, "Feature extraction using convolutional neural network for classifying breast density in mammographic images," *SPIE Medical Imaging*, vol. 10134, 2017.

[19] Pablo Fonseca, Benjamin Castañeda, Ricardo Valenzuela, and Jacques Wainer, *Breast Density Classification with Convolutional Neural Networks*, pp. 101–108, Springer International Publishing, 2017.