2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)
Venice, Italy, April 8-11, 2019

BREAST DENSITY QUANTIFICATION USING WEAKLY ANNOTATED DATASET

Mickael Tardy[‡§] *Bruno Scheffer, MD*[§⋆] *Diana Mateus*[‡]

[‡]Lab. des Sciences du Numérique de Nantes UMR6004, Ecole Centrale Nantes, France
[⋆] Institut de cancérologie de l'Ouest, Nantes, France
[§]Hera-MI, SAS, Nantes, France

ABSTRACT

Breast density is known to be an efficient biomarker for cancer risk, and of particular interest in *early* breast cancer detection, when masses are not yet visible. The quantification of the breast density is difficult due to limitations of mammography imaging, as well as to the ambiguities in defining the limits of the relevant regions. Though inherently a regression task, breast density quantification has been typically approached as a rough classification problem. In this paper, we model the problem of breast density evaluation as an image-wise regression task that seeks to quantify the percentage of fibroglandular tissue. We propose a deep learning method offering a clinically acceptable estimate with low requirements on expert annotations. We also discuss the use of the X-ray acquisition parameters as additional input to the neural network. Our best performing model yields an optimistic mean absolute error around 6.0% of breast density.

***Index Terms*—** mammography, breast density, quantification, deep learning, x-ray, ordered classification.

1. INTRODUCTION

Breast cancer is the most prevalent cancer amongst women and one of the leading causes of death [1]. The chances of recovery go up to 87% if the cancer is detected in early stages [2]. Among other variables, the breast density or mammographic percent density (PD) has been proven to be an efficient biomarker for breast cancer development risk [3], with increasing risk for the denser breasts [4]. Thereby, the breast density evaluation has become mandatory in United States following the Breast Density reporting law and the recommendations of the American College or Radiology.

Percent density (PD) is the ratio of the amount of fibroglandular tissue to the overall breast volume, correlated to the breast dimension and elasticity. Clinical protocols include an approximate classification of the density. For instance, the Breast imaging-reporting and data system (BI-RADS 4th-ed.), defines 4 classes according to the percentage of surface occupied by the dense tissue as follows: Class 1 - $[0\%, 25\%)$, Class 2 - $[25\%, 50\%)$, Class 3 -$[50\%, 75\%)$, Class 4 - $[75\%, 100\%]$ (See Fig. 1 for illustration).

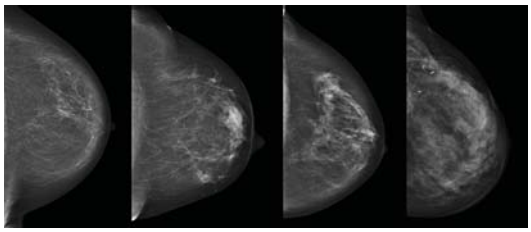

Fig. 1. Left to right, types 1 to 4 according to BI-RADS 4th ed.

In this work, we target a data-driven computer-aided solution for quantification of breast density from Full Field Digital Mammography (FFDM) images. There are several challenges associated to this goal.

First, expert labels are difficult to gather. Clinical protocols only require a rough classification after visual inspection of the image, while the delineation of the relevant tissue in research protocols is subject to a large intra-expert variability [5]. Second, the mammography summarizes the volumetric information in a 2-D projection.

Finally, mammographic acquisition parameters such as exposure time, current, voltage are calibrated to adapt to each breast, and in particular to its density, using automatic exposure control (AEC) [6], and thereby leading to images with different aspect.

We propose a machine learning method to quantify breast density. Given the challenges above, we believe that building a suitable dataset for precise supervised segmentation of the fibroglandular tissue is impractical. Instead, we argue that a finer-grained breast density analysis, closer to the inherent regression nature of the problem, may be enough provide the required support for more personalized treatment. Therefore, we approach the regression task as an ordinal classification problem, and propose to consider 12 classes instead of 4.

Moreover, to cope with the influence of the machine parameters, we propose to combine the image with acquisition parameters and demonstrate an improved performance over the image-only based approach. To the best of our knowledge, we are the first to treat finer grained ordered classification for breast density estimation and to consider the acquisition pa-

Research funding is provided by Hera-MI, SAS and Association Nationale de la Recherche et de la Technologie

rameters model allowing to quantify breast density.

2. RELATED WORK

Multiple contributions have been recently made in the field of mammography image analysis, in particular, by means of deep learning techniques. Amongst them, several works have focused on the detection, localization, segmentation and classification of mass lesions and microcalcifications [7, 8, 9, 10].

While considerable achievements have been shown, mass detection is not suitable for early breast cancer detection task, as masses are not yet detectable. Therefore, more relevant biomarkers, such as breast density, are needed.

Arefan et al. [11] were amongst the first to propose the use of the deep learning techniques for the estimation of the breast density using a classification approach with 3-categories (Fatty, Glandular, Dense) with high accuracy scores. Mohamed et al. [12] deals with 2-categories classification, while and Wu et al. [13] extended the task to 4 classes. In all three cases, such approximate estimations are not precise enough to allow personalized patient treatment.

Li et al. [14] and Wei et al. [15] focus instead on the dense tissue segmentation task, showing promising results with deep learning techniques. We argue, however, that such segmentation approaches are impractical, given the demanding requirements on expert annotations for training and validation.

We propose a different approach to breast density quantification compared to the state of the art [13, 14, 16, 17] introducing an alternative to both, classification and pixel-wise segmentation techniques. Our contributions over prior work include: i) approaching the breast density quantification with a regression model ii) proposing an extended BI-RADS classification system which leads to a higher precision, and iii), considering the acquisition parameters as input-features to our model in order to better cope with x-ray specificities.

3. METHODS

Percent density (i.e. percentage of the fibroblandular tissue) may be specified as:

$$PD = \frac{FT}{V},$$

with FT the amount of the fibroglandular tissue and V the overall volume of the breast. We aim to estimate the PD value for the whole breast by analyzing the whole FFDM image and having access to acquisition parameters. We assume PD varies from 0% to 100%, excluding the skin envelope from the breast volume V. Our goal is to build a model f capable of predicting an estimate of the breast density $\hat{d}_i \in [0, 100]$, given one FFDM image I_i and its acquisition parameters p_i.

$$\hat{d}_i = f(I_i, p_i)$$

To address this problem with a data-driven machine learning approach, we collect a dataset of N images with their parameters and target values, i.e. $\{I_i, p_i, d_i\}_{i=1}^{N}$. Each vector of acquisition parameters contains the following information: voltage (kV), exposure time (ms), tube current (mA), exposure (mAs), entrance Dose (μGy) and retained dose (dGy), as well as compression force (N), angle (deg), breast thickness (mm) and projection area (mm^2).

Given the impracticality of collecting continuous (3D pixel-wise) expert annotations, we propose to extend the current classification standard grid to better address our regression task. First, we use the 4-class BI-RADS grid with a span of 25% per class. Then, we ask the expert radiologist to further grade the images in each class among three subclasses, leading to a 12-class grid with a ~8.5% density span per class (see Fig. 2). Finally, since we use a regression loss on classification labels, our problem ends up being an ordinal classification, which exploits the distance to the target class to guide the learning process.

Fig. 2. 4- and 12-class grid and corresponding breast densities in percentage.

We use a deep neural network to model f, and evaluate the effect of different losses and model choices. We used the Mean Absolute Error (MAE) and Mean Squared Error (MSE) as regression losses. For comparison, categorical cross entropy loss was used as target function for the initial classification task.

Our final model $f_{\mathrm{reg12-metapar}}$ is represented in Fig. 3. It relies on a VGG-like network pre-trained on the 4-class classification problem and fine-tuned to solve the regression task with 12 possible target values. Note that doing a pre-training reduces the amount of additional fine-grid expert annotations required, so we only use extended labels for 25% of the train-set images.

We additionally study four other approaches for comparison: (1) 4-class classification, (2) 4-class regression without parameters, (3) 12-class fine-tuning regression and (4) 4-class

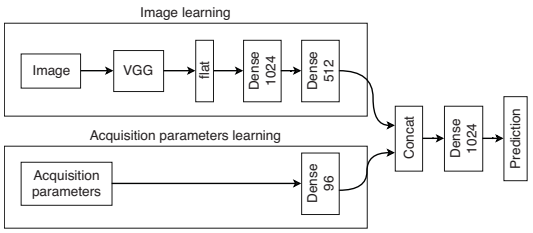

Fig. 3. Fusion model with concatenated image and acquisition parameters branches.

regression with parameters.

As a proof of concept, we add the information from the acquisition parameters as features of a parallel processing branch. Both the image and the acquisition parameters are passed through dense layers before being concatenated. The output of the concatenation was fed to a large dense layer before being inserted into the final regression layer.

The motivation for the baseline 4-class classification f_{class4} is the practical availability of the ground truth since the classes are given systematically by the radiologist in the clinical routine.

To highlight the advantages of training with a regression loss, we also address the coarse 4-class grid with f_{reg4} using the same training dataset and network from f_{class4}, except for replacing the top classification layers of f_{class4} with a regression layer. Other settings of the network remained identical.

Finally, we rely on the pre-trained classification model f_{class4} to resolve the finer-grid regression task f_{reg12}. As before, the top classification layer was replaced with a regression layer, while all other settings remained identical. For fine-tuning, we used again the smaller dataset containing more precise expert-annotations with 12 target values.

3.1. Network implementation and optimisation details

We used the same neural network for all methods: a VGG-like [18] architecture known to be efficient for mammography [10]. The network is composed of six blocks of two convolutional layers each with number of filters as follows: $32/32$ - $64/64$ - $128/128$ - $256/256$ - $256/256$ - $512/512$. The kernels of all convolutional filters are set to 3×3. After each convolutional layer with ReLU activation we perform batch normalization. At the end of each block, we have a max pooling layer with pool size of 2×2 and strides of 2×2. Finally, after the convolutional network we added two dense layers. For the classification the result of dense layers has been fed to a 4-classes prediction layer. As for regression network, we only replaced the prediction layer by 1-class regression layer.

The architecture choice for the acquisition parameters branch followed an independent set of experiments to identify the best performing regression model based on the parameters only. The most efficient retained model was the one wide dense layer network.

4. EXPERIMENTAL VALIDATION

4.1. Dataset

Publicly available datasets, such as Digital Database for Screening Mammography (DDSM) or INBreast do not provide the required fine density annotations neither acquisition parameters. Thus, we collected an in-house dataset composed of 1602 images from 283 patients and 434 different exams. It includes images from both the cranocaudal (CC) and the mediolateral oblique (MLO) views. The dataset is split in training and test sets, which remains the same throughout the experiments. The train set contains 1232 images (70%) and the test set the remaining 370 images (30%). The split ensures that different views of the same breast are kept in the same subset.

All images have been labeled by an expert breast imaging radiologist using the 4th edition of BI-RADS classification system. Moreover, in order to train the system efficiently for the regression task, a subset of 282 training images as well as the total number of test images (370) were annotated with the extended 12-class system (see Fig. 2). In order to diminish eventual bias associated with the lack of multiple expert opinion, the same data were evaluated by the expert three times under different conditions (e.g. on different workstations). The final target value used for each image is the majority voting value(i.e. 2 out of 3).

We note that the use of the more precise ground truth such as MRI may be beneficial for the model performance, however the breast MRI data are usually less common in clinical practice so their collection needs bigger efforts.

The images have been pre-processed as follows: (1) crop to remove empty background pixels, (2) re-scale the image setting its longer side to 256 pixels, (3) if necessary, flip the image horizontally to systematically align the pectoral muscle to the left and (4) pad the shortest edge of the image with empty pixels to obtain a squared 256×256 image.

4.2. Network training

For all models, the training has been performed per epoch on the entire training dataset. In total, 1000 epochs were let to run for each model. The validation has been performed on the entire test dataset after every 25 epoch. In order to prevent any unnatural image modification we did not use any augmentation.

While training the model for classification, we applied class weighing in order to compensate for imbalance. Also, for the final model, we pre-trained separately both the image and the acquisition-parameters branches to favor the overall convergence of the model.

4.3. Validation details

We evaluated classification and regression performance of the 5 studied models with the objective to prove the advantages of the regression approach over the classification, as well as the benefits of the proposed network and training modifications.

The tests were performed systematically on the same dataset of 370 images, which contains both 4 and 12 class annotations. The classification performances were collected on the 4-classes grid, while the regression performances were systematically compared against 12-class annotations.

For the assessment of classification performance we used the following metrics: accuracy, precision, recall, F_1-score

Table 1. Classification performances of the studied models

Model	Metrics				
	Accuracy	***Precision***	***Recall***	***F_1-score***	***Cohen kappa***
f_{class4}	0.741 CI: 0.729 - 0.752	0.749 CI: 0.739 - 0.758	0.741 CI: 0.729 - 0.752	0.738 CI: 0.726 - 0.750	0.850 CI: 0.838 - 0.863
f_{reg4}	0.759 CI: 0.740 - 0.779	0.780 CI: 0.767 - 0.793	0.759 CI: 0.740 - 0.779	0.762 CI: 0.741 - 0.782	0.879 CI: 0.858 - 0.900
f_{reg12}	0.764 CI: 0.757 - 0.771	0.782 CI: 0.778 - 0.786	0.764 CI: 0.757 - 0.771	0.766 CI: 0.760 - 0.773	0.891 CI: 0.887 - 0.896
$f_{\text{reg4-metapar}}$ (fixed weights)	0.784 CI: 0.782 - 0.786	0.800 CI: 0.798 - 0.801	0.784 CI: 0.782 - 0.786	0.787 CI: 0.785 - 0.788	0.899 CI: 0.898 - 0.901
$f_{\text{reg12-metapar}}$ (fixed weights)	**0.796 CI: 0.792 - 0.800**	**0.811 CI: 0.808 - 0.814**	**0.796 CI: 0.792 - 0.800**	**0.797 CI: 0.793 - 0.800**	**0.906 CI: 0.904 - 0.908**

Table 2. Regression performances of the studied models

Model	Metrics		
	MAE	***MxAE***	***C-index***
f_{class4}	8.873 CI: 8.510 - 9.236	68.590 CI: 65.250 - 71.930	0.809 CI: 0.802 - 0.816
f_{reg4}	7.520 CI: 6.743 - 8.298	40.640 CI: 36.036 - 45.244	0.826 CI: 0.821 - 0.832
f_{reg12}	6.545 CI: 6.379 - 6.712	31.964 CI: 31.214 - 32.714	0.820 CI: 0.815 - 0.824
$f_{\text{reg4-metapar}}$ (fixed weights)	6.434 CI: 6.397 - 6.471	30.274 CI: 29.293 - 31.256	0.831 CI: 0.827 - 0.835
$f_{\text{reg12-metapar}}$ (fixed weights)	**6.092 CI: 6.030 - 6.154**	**28.113 CI: 27.541 - 28.685**	**0.843 CI: 0.840 - 0.847**

Table 3. Confusion matrix of Classification model f_{class4}

Truth	Predictions			
	1	***2***	***3***	***4***
1	41	11	0	0
2	8	80	4	0
3	2	23	41	20
4	0	4	11	125

Table 4. Confusion matrix of Fine-trained regression model $f_{\text{reg4-metapar}}$

Truth	Predictions			
	1	***2***	***3***	***4***
1	34	18	0	0
2	2	70	20	0
3	0	5	68	13
4	0	0	10	130

and Cohen's kappa [19] comparing the agreement of the algorithm with the expert. For the regression task we relied essentially on the mean absolute error (MAE). We also report the maximum absolute error (MxAE), which is the maximum value amongst all the absolute errors on test dataset.

The interest of the MxAE is to highlight the maximum span of the misclassification that may be critical in clinical application. Moreover, we also compared the concordance index to evaluate how well the predictions of the different models respect the class orders.

We report the mean and confidence intervals (CI) of the metrics computed at several time points during the training, starting from epoch 200, where our models started to converge. The CIs are all calculated with 0.95 confidence.

5. RESULTS

In Tab. 1, we report the classification performance. The three image models have comparable results, while $f_{\text{reg12-metapar}}$ presents an advantage on all metrics and a 5% accuracy increase. The disadvantages of the straightforward classification approach are highlighted by the MxAE (see Tab. 2) and the confusion matrice in Tab. 3: they lead to critical misclassification errors ($> 50\%$). In this sense, the regression task is safer, with the f_{reg4} notably decreases MxAE from 68.6% to 40.6%, and our f_{reg12} further decreasing such errors to 31.96% (see Tab. 2). While we observe worse performance on 1st and 2nd classes, we note the benefit of smaller error span, as well as lower density underestimation (see Tab. 4).

Adding the meta-parameters to our model yielded an additional increase in performance. We note the benefit of $f_{\text{reg12-metapar}}$ compared to $f_{\text{reg4-metapar}}$. We observe a gain in MxAE (28.1%) and an increase in accuracy (0.796) compared to f_{reg12}.

We experimented with fine-tuning of the entire model versus the dense layers only. In case of f_{reg12} we observed an eventual leak of MxAE (i.e. $MxAE > 75\%$) when the convolutional layers weights were trainable.

During our work we were able to estimate an intra-reader kappa of $k = 0.93$, having evaluated the dataset multiple times. Remarkably, our best performing model yields similar agreement ($k = 0.906$) to the expert, showing the capability of the system to reproduce reader's behavior.

6. DISCUSSION AND CONCLUSION

We have studied the problem of breast density quantification with the limitations of clinically available annotations. We evaluated classification and regression approaches using fine tuning on a small but fine-grained dataset. This allowed us to obtain a performant model with an accuracy of (0.796 CI: 0.792 - 0.800) and mean absolute error of (6.092 CI: 6.030 - 6.154). Our method suits well two tasks, regression for quantitive and classification for qualitative analyses. The MxAE error is brought to 28.1%, which is comparable to one class step in the BI-RADS 4ed grid. We observe increase of the performance with both, 12-class annotations and inclusion of meta-parameters. The 4-class model with meta-parameters is the runner-up proving the usefullness of acquisition data.

Our solution has several clinical applications. First, it offers a clinically acceptable estimation of breast density, which is in increasing demand. Second, the proposed fine breast density quantification provides additional guidance for personalized healthcare. Third, the system may help radiologist in daily routine by prioritizing cases accessing more complex cases at the moments of higher awareness. Lastly, it contributes further to the consideration of breast density as biomarker.

Future research includes collecting annotations from multiple reviewers and studying other means to include the acquisition parameters.

References

[1] Rebecca L. Siegel, Kimberly D. Miller, and Ahmedin Jemal, "Cancer statistics, 2018," *CA: A Cancer Journal for Clinicians*, vol. 68, no. 1, pp. 7–30, jan 2018.

[2] Béatrice Lauby-Secretan, Chiara Scoccianti, Dana Loomis, Lamia Benbrahim-Tallaa, Véronique Bouvard, Franca Bianchini, and Kurt Straif, "Breast-Cancer Screening Viewpoint of the IARC Working Group," *New England Journal of Medicine*, vol. 372, no. 24, pp. 2353–2358, jun 2015.

[3] Celine M. Vachon, Carla H. van Gils, Thomas A. Sellers, Karthik Ghosh, Sandhya Pruthi, Kathleen R. Brandt, and V. Shane Pankratz, "Mammographic density, breast cancer risk and risk prediction," *Breast Cancer Research*, vol. 9, no. 6, pp. 217, dec 2007.

[4] Norman F. Boyd, Helen Guo, Lisa J. Martin, Limei Sun, Jennifer Stone, Eve Fishell, Roberta A. Jong, Greg Hislop, Anna Chiarelli, Salomon Minkin, and Martin J. Yaffe, "Mammographic Density and the Risk and Detection of Breast Cancer," *New England Journal of Medicine*, vol. 356, no. 3, pp. 227–236, 2007.

[5] Afsaneh Alikhassi, Hamed Esmaili Gourabi, and Masoud Baikpour, "Comparison of inter- and intra-observer variability of breast density assessments using the fourth and fifth editions of Breast Imaging Reporting and Data System," *European Journal of Radiology Open*, vol. 5, pp. 67–72, 2018.

[6] S Sterling, "Automatic exposure control: a primer," *Radiol Technol*, vol. 59, no. 5, pp. 421–427, 1988.

[7] Gustavo Carneiro, Jacinto Nascimento, and Andrew P. Bradley, "Automated Analysis of Unregistered Multi-View Mammograms with Deep Learning," *IEEE Transactions on Medical Imaging*, vol. 36, no. 11, pp. 2355–2365, nov 2017.

[8] Neeraj Dhungel, Gustavo Carneiro, and Andrew P. Bradley, "A deep learning approach for the analysis of masses in mammograms with minimal user intervention," *Medical Image Analysis*, vol. 37, pp. 114–128, 2017.

[9] Thijs Kooi, Geert Litjens, Bram van Ginneken, Albert Gubern-Mérida, Clara I. Sánchez, Ritse Mann, Ard den Heeten, and Nico Karssemeijer, "Large scale deep learning for computer aided detection of mammographic lesions," *Medical Image Analysis*, vol. 35, pp. 303–312, jan 2017.

[10] Dezso Ribli, Anna Horváth, Zsuzsa Unger, Péter Pollner, and István Csabai, "Detecting and classifying lesions in mammograms with Deep Learning," *Scientific Reports*, vol. 8, no. 1, pp. 4165, dec 2018.

[11] D. Arefan, A. Talebpour, N. Ahmadinejhad, and A. Kamali Asl, "Automatic breast density classification using neural network," *Journal of Instrumentation*, vol. 10, no. 12, 2015.

[12] Aly A. Mohamed, Wendie A. Berg, Hong Peng, Yahong Luo, Rachel C. Jankowitz, and Shandong Wu, "A deep learning method for classifying mammographic breast density categories," *Medical Physics*, vol. 45, no. 1, pp. 314–321, jan 2018.

[13] Nan Wu, Krzysztof J. Geras, Yiqiu Shen, Jingyi Su, S. Gene Kim, Eric Kim, Stacey Wolfson, Linda Moy, and Kyunghyun Cho, "Breast density classification with deep convolutional neural networks," Tech. Rep., 2017.

[14] Songfeng Li, Jun Wei, Heang Ping Chan, Mark A. Helvie, Marilyn A. Roubidoux, Yao Lu, Chuan Zhou, Lubomir M. Hadjiiski, and Ravi K. Samala, "Computer-aided assessment of breast density: Comparison of supervised deep learning and feature-based statistical learning," *Physics in Medicine and Biology*, vol. 63, no. 2, pp. 025005, jan 2018.

[15] J. Wei, S. Li, H.-P. Chan, M.A. Helvie, M.A. Roubidoux, Y. Lu, C. Zhou, L. Hadjiiski, and R.K. Samala, "Deep convolutional neural network for mammographic density segmentation," in *Progress in Biomedical Optics and Imaging - Proceedings of SPIE*, Kensaku Mori and Nicholas Petrick, Eds. feb 2018, vol. 10575, p. 126, SPIE.

[16] Brad M. Keller, Jinbo Chen, Dania Daye, Emily F. Conant, and Despina Kontos, "Preliminary evaluation of the publicly available Laboratory for Breast Radiodensity Assessment (LIBRA) software tool: Comparison of fully automated area and volumetric density measures in a case-control study with digital mammography," *Breast Cancer Research*, vol. 17, no. 1, pp. 117, 2015.

[17] Aimilia Gastounioti, Emily F. Conant, and Despina Kontos, "Beyond breast density: A review on the advancing role of parenchymal texture analysis in breast cancer risk assessment," 2016.

[18] Karen Simonyan and Andrew Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," Tech. Rep., 2014.

[19] Jacob Cohen, "A Coefficient of Agreement for Nominal Scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, apr 1960.