Photometric Transformer Networks and Label Adjustment for Breast Density Prediction

Jaehwan Lee
Lunit Inc.
Seoul, Republic of Korea
jhlee@lunit.io

Donggeun Yoo
Lunit Inc.
Seoul, Republic of Korea
dgyoo@lunit.io

Hyo-Eun Kim
Lunit Inc.
Seoul, Republic of Korea
hekim@lunit.io

Abstract

Grading breast density is highly sensitive to normalization settings of digital mammogram as the density is tightly correlated with the distribution of pixel intensity. Also, the grade varies with readers due to uncertain grading criteria. These issues are inherent in the density assessment of digital mammography. They are problematic when designing a computer-aided prediction model for breast density and become worse if the data comes from multiple sites. In this paper, we proposed two novel deep learning techniques for breast density prediction: 1) photometric transformation which adaptively normalizes the input mammograms, and 2) label distillation which adjusts the label by using its output prediction. The photometric transformer network predicts optimal parameters for photometric transformation on the fly, learned jointly with the main prediction network. The label distillation, a type of pseudo-label techniques, is intended to mitigate the grading variation. We experimentally showed that the proposed methods are beneficial in terms of breast density prediction, resulting in significant performance improvement compared to various previous approaches.

1. Introduction

Breasts can be categorized as *dense* or *fatty* by the portion of parenchyma in the breasts. A fatty breast indicates that the breast is mostly composed of fat tissue, whereas a dense breast has more dense tissue that shows dense parenchymal patterns on mammograms. Readers should be more careful when dealing with mammograms with dense parenchymal pattern since suspicious malignant lesions can be hidden, resulting to a false-negative [6]. Also, it has been reported that a dense breast has a higher risk of breast cancer than average [1]. For this reason, BI-RADS [11], which is a standard protocol for breast imaging, guides the interpreting readers to report density category as an essential field of case reports form(CRF). In BI-RADS taxonomy, breast density is categorized into four grades: *a*, *b*, *c*, *d*, meaning "almost entirely fatty", "scattered areas of fibro-glandular tissue", "heterogeneously dense", and "extremely dense", respectively.

Based on the collected mammograms and their density categories in CRFs, it is straight-forward to regard a density prediction task as classification. However, breast density prediction is not a typical classification task. The BI-RADS criteria for breast density are 1) the portion of parenchyma within a breast, which is discretization of the continual score, and 2) specific dense parenchyma pattern in part of the image, determined by the reader. Thus, the density labels in a training dataset will have inter-readers biases.

Intensity normalization of mammograms is an important factor when grading the breast density, since the mammographic parenchymal pattern is highly correlated with the pixel intensity. However, intensity distribution of the parenchyma and the fat tissue varies according to different vendors of imaging devices as well as different hospitals. To compensate these variations, readers often manually adjust the contrast of each mammogram to determine the grade properly.

In this paper, we propose two methods that tackle the problems caused by the normalization and inter-reader grading variance. The first method is a learnable normalization module, called photometric transformer network (PTN), that predicts normalization parameters of input mammogram. It is seamless to main prediction network so that optimal normalization and density grade can be learned jointly. The second one is a label distillation method, which is a type of pseudo-label technique, taking the grading variation into consideration.

Our test shows that proposed two methods help to improve performance, especially in multi-site configurations. Our final model outperforms other public-available previous models in a test set with neutral configurations. Experimental results show that the proposed method improves the accuracy and $dAUC$ (a novel evaluation metric of the den-

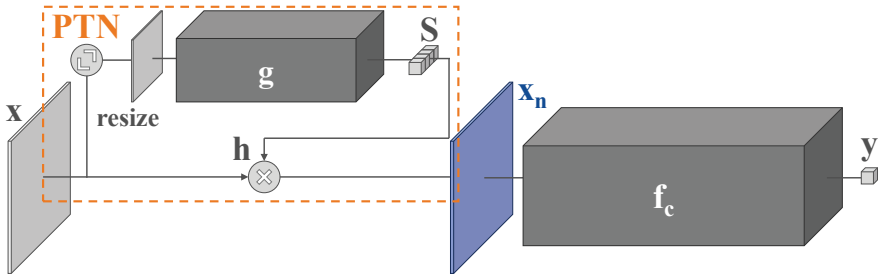


Figure 1. An overview of our model architecture. Before a mammogram is input to the classifier network, the intensity distribution of the mammogram is changed with S which are outputs of transformer network.

sity grading) from 55% to 79% and from 0.9204 to 0.9663, respectively. The proposed method also outperforms previous state-of-the-arts based on an evaluation on external test data, which is collected from a separate institution for fair comparison between similar approaches.

2. Related works

With the drastic advance of deep learning, breast density prediction based on deep neural networks has also been introduced recently. [4] applied the unsupervised feature learning based on auto-encoder to predict the breast density. [7, 9, 14] employed convolutional neural networks (CNNs) that is learned with a cross-entropy loss for breast density prediction. Motivated by these approaches, we also cast the breast density prediction as a CNN-based classification task, but address the two practical problems caused by multi-site configuration.

From the perspective of dynamic estimation of the parameters which are appropriate for a target task, our PTN is similar to the spatial transformer network [3]. Spatial transformer network predicts appropriate geometric transformation parameters, while our PTN tries to find a set of photometric transformation parameters that is optimal for breast density prediction.

The proposed label distillation is motivated by pseudo-labeling techniques, devised especially for handling label noise [8, 12]. In [8], an auxiliary network trained with small clean examples were used to predict pseudo-labels, in addition to the main network trained with large examples with given pseudo-labels. Similarly in [12], a sub-network jointly optimized with a main network tries to find appropriate pseudo-labels. Our approach is distinct from [8, 12], in that pseudo-labels are given to only selected samples and applied in iterative ways to prevent distillation of model bias.

3. Methods

A density estimator f is a neural network that predicts breast density $y \in \{a, b, c, d\}$ from an input mammogram $x \in \mathbb{R}^{H \times W}$. The input x is normalized by the PTN denoted by f_n, and the classifier f_c estimates density $\hat{y}$ from the normalized input as

$$\hat{y} = f_c(f_n(x; \theta_n); \theta_c), \tag{1}$$

where f_n and f_c are parameterized by θ_n and θ_c, respectively. Our goal is to learn parameters $\theta = \theta_n \cup \theta_c$ with our dataset $D = \{(x_i, y_i) \,|\, i = 1, \cdots, N\}$.

$$\theta^* = \arg\min_{\theta} \frac{1}{N} \sum_{(x,y) \in D} \mathcal{L}(\hat{y}, y) \tag{2}$$

where $\mathcal{L}$ is the loss function. To successfully estimates θ, we propose the photometric transformer module f_n in Section 3.1, and a distillation method to handle label grading variance problem in Section 3.2.

3.1. Photometric transformer networks

The f_n normalizes an input x by a function h. The function h is determined by a parameter set S, and the parameter set S is predicted by a CNN g from the input x. For a pixel intensity $x(i, j)$ at a location (i, j), it can be expressed as

$$x_n(i, j) = h(x(i, j), S) \quad \text{where} \quad S = g(x; \theta_n), \tag{3}$$

and is illustrated in the left of Figure 1.

We introduce the function h that works well in breast density prediction. Let us assume the intensity range of interests is $[u, v)$[1]. We split the range into K sub-intervals, giving $T_k = [u + t(k-1), u + tk)$ where $t = (v - u)/K$ and $k = 1, \cdots, K$. Then, the function h is defined as

$$h(x(i,j), S) = \begin{cases} u + s_0(x(i,j) - u) \\ \qquad \text{if } x(i,j) \in (-\infty, u) \\ u + \sum_{l=1}^{k-1} s_l t + s_k(x(i,j) - \min(T_k)) \\ \qquad \text{if } x(i,j) \in T_k \\ u + \sum_{l=1}^{k} s_l t + s_{K+1}(x(i,j) - v) \\ \qquad \text{if } x(i,j) \in [v, \infty) \end{cases} \tag{4}$$

[1]Generally, it is determined by window center & width value in standard DICOM protocols.

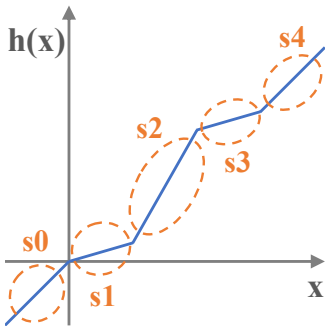


Figure 2. Example graph of the proposed function h. The domain of h is divided into $K+1$ intervals. The slope of each interval are defined with the parameters generated by the network g.

where S is $\{s_0, \cdots, s_{K+1}\}$, and $\min(T_k)$ is a minimum value of an interval T_k. Figure 2 is illustrating this function h. Each component of S can be interpreted as a slope of the corresponding line segment.

The function h is continuous but can fluctuate if a part of S is negative. To make h be an increasing function, we add a hinge regularization term to the cross entropy loss $\mathcal{L}_{\text{CE}}$. The loss function in Equation (2) is finally defined as

$$\mathcal{L}(\hat{y}, y) = \mathcal{L}_{\text{CE}}(\hat{y}, y) + \lambda \cdot \sum_{i=0}^{k+1} \min(-s_i, -\epsilon) + \epsilon \qquad (5)$$

where ϵ is a small positive constant and λ is a scaling constant. We have empirically found that adding this regularization term yields better performance.

3.2. Label distillation

We propose a novel approach to train a model in the situation that having inter-reader (i.e., inter-labeler) grading variance. The dataset D is divided by D_s which is labeled by a single reader and the rest D_r. D_s is the set having small samples but free from the variance, while D_r is the large set suffering from the variance from arbitrary readers.

The proposed method consists of three steps. In the first stage, training is performed with D_r, as a typical machine learning algorithm does. In the second stage, the model is fine-tuned with D_s. This transfer learning strategy is for reducing the inter-reader variance while utilizing the general mid-level image representation learned with a large number of samples in D_r.

In the third stage, labels in D_r are refined to pseudo-labels generated by the model produced by the second stage. The new labels $\{\hat{y}|x \in D_r\}$ have the grading criteria that is more similar to that of the single reader labels in D_s. This procedure from the first to the third stage is repeated until the model converges.

Pseudo-labeling distillates not only the grading criterion used in labeling D_s, but also the knowledge of the model as argued in [2]. Unfortunately, hard samples that wrongly labeled by the model fitted to D_s distillates inaccurate knowledge.

Our method tries to filter out hard samples to prevent conveying the inaccurate knowledge. To this end, we measure a divergence $\text{KL}(y, \hat{y})$ between a one-hot encoded label y and prediction $\hat{y}$ for each sample $x \in D_r$. We empirically found that a hard sample is prone to have a relatively small divergence value rather than the sample with inaccurate label. We select top γ-percent samples of $\text{KL}(y, \hat{y})$ to filter out the hard samples. After that, for each of the selected samples, we update y with $\hat{y}$ by blending operation as

$$y := \alpha y + (1 - \alpha) \cdot \hat{y}, \qquad (6)$$

where α is a constant blending factor. We then continue training with the updated D_r. This procedure is repeated until the model converges.

The whole procedure of the proposed method is concretely described in Algorithm 1.

Algorithm 1 Label distillation

$\theta := \arg\min_\theta \frac{1}{N} \sum_{(x,y)\in D} \mathcal{L}(f(x;\theta), y)$
Split D into D_s and D_r
while not converged **do**
  $\theta := \arg\min_\theta \frac{1}{N} \sum_{(x,y)\in D_s} \mathcal{L}(f(x;\theta), y)$
  for $(x, y) \in D_r$ **do**
    if KLD$(y, f(x;\theta))$ is top r% **then**
      $y := \alpha y + (1 - \alpha) \cdot f(x;\theta)$
    end if
  end for
  $D_{\text{train}} := D_s \cup D_r$
  $\theta := \arg\min_\theta \frac{1}{N} \sum_{(x,y)\in D_{\text{train}}} \mathcal{L}(f(x;\theta), y)$
end while
return θ

Table 1. Dataset configurations

Datasets\Grades	a	b	c	d	Total
Training set D_r	1,395	6,905	33,282	4,773	46,355
Training set D_s	72	391	428	255	1,146
Validation set	78	373	421	275	1,147
Test set	9	280	455	242	986
External test set	852	3,130	3,634	590	8,206

4. Evaluation

4.1. Experimental setup

4.1.1 Datasets

We have collected 48,648 cases of Asian women from 5 separate hospitals from South Korea. Each case comprises four mammograms with different views of a left CC, a left

MLO, a right CC, and a right MLO. We also select approximately 5% samples to refine labels by a single reader, i.e., a radiologist who is a breast specialist. Half of the 5% samples are used for D_s, and the rest for a validation set. As an in-house test set, we have collected 986 cases from another institution from South Korea. The same radiologist has labeled this test set. To fairly compare ours with other method, we have collected another test set(external test set), which comprises 8,206 cases, from a large hospital in the US. We have extracted the density grade for each case from CRF field, and use it as a label. Table 1 summarizes our datasets.

4.1.2 Baseline

For classifier f_c, we adopt ResNet-18 and make it produce a 4-dimensional softmax output. We used SGD optimizer and the learning rate is set to 0.1 in training. The model takes a single mammogram as input, and four predictions from four views are averaged to a case-level prediction. We decode mammograms by the window center and width embedded in DICOM protocol.

To check the sanity of our baseline networks, the baseline is compared with other neural-network methods, [7] and [14]. Roundabout way have to be used for comparison, since all reported scores in other works are obtained with different configurations and private datasets. The training split and the test set split collected in the same site in [7] and [14], while our training split and test split are from the other sites. To keep the configuration of the experiment as same as possible, we have followed same experiment settings proposed in [7] and [14] for this experiment. We split our main dataset to three parts, D_r for training, D_s for validation and testing set is validation set in original split.

Two metrics are tracked same as [7] and [14]: 4-class accuracy and 2-class(fatty vs. dense) accuracy. Class-wise averaged accuracy is used in this sanity check. 4-class accuracy and 2-class accuracy have reported approximately **77**% and **87**% in their papers (in same), while our baseline reports **74**% and **89**%. Our baseline model is inferior in 4-class accuracy than other models, but it is superior in 2-class accuracy. Interestingly, [7] and [14] also reported almost same scores each other in their in-house test set. Putting the above results together, we have concluded that our model has almost similar accuracy to other works. It means if generalization for inter-reader variance is not considered, even they trained with all different datasets and different hyper-parameters, the capability of classifying breast density scores are almost the same. Note that the above accuracy score is class-wised accuracy, while all reported accuracy scores through the paper are instance-wise accuracy. This is because make metric comparable to other works.

4.1.3 Metrics

We use the 4-way classification accuracy, as it has been the common metric for previous works. Unfortunately, class-averaged accuracy scores may be inaccurate in our test set since our test set suffers class-imbalance problem, existing only 9 samples with category a. For example, a sample with category a contributes to accuracy $455/9 = 50.56$ times more than the another sample that category is c in class-averaged accuracy metrics. Instead of the class-averaged accuracy, the instance-wise average accuracy score is used to relax this problem.

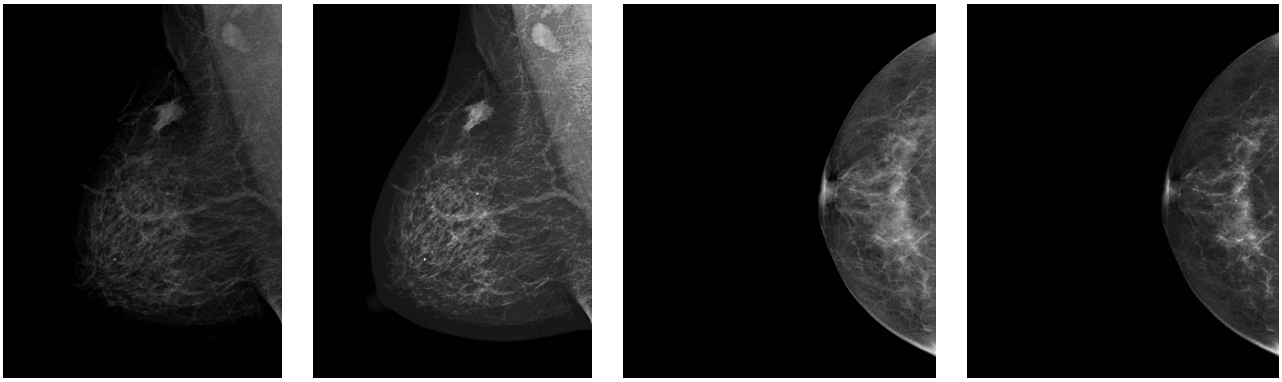
Moreover, the accuracy metric itself is also inaccurate when it takes into accounts inter-reader variance problem. This is because breast density prediction is not a typical classification task. In whatever ways of grading criterion of choosing a discrete category, the sample that is vague to classify between two values, since the grade of the density is discretization of a continual density score that actual physical quantity is a portion of parenchyma in a breast.

In addition to this issue, there exists a relation between labels in breast density, where accuracy metric more inaccurate. For example, the grade a is closer to b, rather than c and d.

To take these issues into account, we propose a new metric, called density-AUC(*dAUC*), which is stands for breast density estimation algorithms. This metric is the aggregation of AUC scores between the density predictions from a model and binarized breast density categories. The labels in AUC should be in forms of binary domain(negative or positive), so the breast density labels in $y \in \{a, b, c, d\}$ are split into three ways: [a vs. b, c, d], [a, b vs. c, d], and [a, b, c vs. d]. In results, we can obtain 3 different labels set for a given dataset, or 3 sub-problems for *dAUC*. Samples having left-side breast density categories are assigned to negative(0), and the ones belongs to right-side are assigned to the positive value(1).

Meanwhile, in addition to label binarization, the predictions of the network needs to be reduced for each sub-problem, since the prediction scores in AUC should be in forms of single real value score. The format of outputs in the proposed model is a vector having length 4, each component represent the probability of each class(a, b, c, d). We take an average of probabilities of each positive categories in each sub-problems. For instance, when we measure an AUC score of [a, b vs. c, d], a sample score is defined as $\hat{y_c}$+$\hat{y_d}$ where $\hat{y_c}$ and $\hat{y_d}$ are the two elements of softmax output $\hat{y}$. This metric satisfy our assumption that defined implicitly – lower value for fatty breast and a higher value for dense breast. The final *dAUC* score is calculated by averaging the three sub-problems.

Note that *dAUC* is just considered as a complement of accuracy metric, not a substitution. Accuracy metric is also tracked as an important metrics. Producing density score



(1) Input (2) Normalized (3) Input (4) Normalized

Figure 3. Original mammograms (1), (3) with the density *a* and *d* are normalized to (2), (4) by PTN. For both samples, labels are corrected to *b*.

with obeying radiologist's convention is also important in practical use.

4.2. Results and analysis

4.2.1 Photometric transformer networks

We use 6 convolution layers for transformer networks and set K to 10. Each mammogram is resized to one-third of the original size. We also use the instance normalization [13] for each of the convolutions, rather than batch normalization. We empirically found that the result image that transformed by PTN is more well-normalized when PTN contains instance normalization.

The standard deviation of image-level pixel mean values in the validation set is 0.6850. Once PTN normalizes the validation images, the standard deviation is significantly reduced to 0.2249. This verified that PTN suppresses inter-image intensity variations and make the image intensities more consistent. Fig. 3 shows some of normalization examples.

The upper part of Table 2 shows the results of normalization methods. CLAHE [10] is selected for representative of static normalization approach. CLAHE improves the baseline, but PTN shows better performance.

4.2.2 Label distillation

For label distillation, we choose the best PTN model as our baseline. The pre-trained parameters of the first two layers are fixed, and the rest is trained with a learning rate of 0.01. We set α and γ as 0.5 and 0.25, respectively. The lower part of Table 2 shows the results.

[2]3 median results of above PTN results, in perspective of validation accuracy.

[3]It produces the percent of density value directly, thus the only dAUC is reported.

The hard-labeling, which directly uses predictions of PTN as pseudo labels, shows higher accuracy score than the baseline but lower dAUC score. [12] is another approach, which uses soft pseudo labels. To make [12] fairly comparable to our method, we fine-tune with D_s at each epoch, before giving pseudo labels. This trial improves both accuracy and dAUC scores, however, the gains are not significant. In contrast, our label distillation method yields clear performance gains compared to the baseline.

4.3. Comparison with others

We conduct another experiment with different settings to compare our models to other works fairly. Instead of test set used in 4.2, another external test set is used for this experiment. (See 4.1.1 for details.) Two of our models are evaluated: baseline, and the proposed PTN and label distillation applied model. Our model is selected by the median value of accuracy score in the in-house validation set, among five trials of experiment. As external algorithms, LIBRA [5], an open-source density predictor, and some other previous works [7, 14] who have opened their model parameters in public are selected.

The results are shown in Table 3. Although our model is trained with the data consists of different race, our best model achieves the best performance with large margins in all metrics.

5. Conclusion

In this paper, we have proposed two methods for breast density problem: PTN and label distillation. These two methods can resolve input and label issues in the breast density prediction task, respectively. For further research, strict validation of dAUC metric how it is suitable for breast density tasks is needed. Additionally, our approach should be looked in broad views, and applied for various medical imaging problems, since it is not limited to a specific task.

Table 2. Breast density estimation performance comparison between methods. The mean and standard deviation of 5 trials are reported.

	Validation		Test	
Methods	Accuracy	dAUC	Accuracy	dAUC
Baseline	.7015(.0179)	.9595(.0153)	.5452(.1078)	.9204(.0207)
CLAHE [10]	.7374(.0291)	.9654(.0154)	.7163(.0341)	.9357(.0128)
PTN	.7479(.0229)	.9755(.0013)	.7509(.0103)	.9518(.0079)
PTN[2]	.7512(.0109)	.9757(.0014)	.7431(.0046)	.9470(.0045)
PTN + hard labeling	.7367(.0150)	.9715(.0026)	.7671(.0039)	.9392(.0155)
PTN + [12]	.7428(.0126)	.9745(.0018)	.7650(.0128)	.9482(.0067)
PTN + [12] in D_s	.7576(.0118)	.9776(.0015)	.7743(.0029)	.9442(.0029)
PTN + label distillation	.8073(.0043)	.9808(.0009)	**.7941**(.0060)	**.9663**(.0033)

Table 3. Breast density estimation performance comparison between methods on the external test set.

	Accuracy	dAUC
LIBRA[3]	-	.8877
[7]	.5860	.9275
[14]	.5419	.8424
Our baseline	.4246	.9185
Ours	**.7257**	**.9481**

References

[1] N. F. Boyd, H. Guo, L. J. Martin, L. Sun, J. Stone, E. Fishell, R. A. Jong, G. Hislop, A. Chiarelli, S. Minkin, and M. J. Yaffe. Mammographic Density and the Risk and Detection of Breast Cancer. *New England Journal of Medicine*, 356(3):227–236, jan 2007.

[2] G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*, 2015.

[3] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu. Spatial Transformer Networks. In *Advances in Neural Information Processing Systems 28 (NIPS)*, pages 2017–2025, jun 2015.

[4] M. Kallenberg, K. Petersen, M. Nielsen, A. Y. Ng, P. Diao, C. Igel, C. M. Vachon, K. Holland, R. R. Winkel, N. Karssemeijer, and M. Lillholm. Unsupervised Deep Learning Applied to Breast Density Segmentation and Mammographic Risk Scoring. *IEEE Transactions on Medical Imaging*, 35(5):1322–1331, may 2016.

[5] B. M. Keller, D. L. Nathan, Y. Wang, Y. Zheng, J. C. Gee, E. F. Conant, and D. Kontos. Estimation of breast percent density in raw and processed full field digital mammography images via adaptive fuzzy c-means clustering and support vector machine segmentation. *Medical Physics*, 39(8):4903–17, aug 2012.

[6] K. Kerlikowske, D. Grady, J. Barclay, E. A. Sickles, and V. Ernster. Effect of Age, Breast Density, and Family History on the Sensitivity of First Screening Mammography. *The Journal of the American Medical Association*, 276(1):33, jul 1996.

[7] C. D. Lehman, A. Yala, T. Schuster, B. Dontchos, M. Bahl, K. Swanson, and R. Barzilay. Mammographic Breast Density Assessment Using Deep Learning: Clinical Implementation. *Radiology*, 290(1):52–58, jan 2019.

[8] Y. Li, J. Yang, Y. Song, L. Cao, J. Luo, and L. J. Li. Learning from Noisy Labels with Distillation. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1928–1936, oct 2017.

[9] A. A. Mohamed, W. A. Berg, H. Peng, Y. Luo, R. C. Jankowitz, and S. Wu. A deep learning method for classifying mammographic breast density categories. *Medical Physics*, 45(1):314–321, jan 2018.

[10] S. M. Pizer, E. P. Amburn, J. D. Austin, R. Cromartie, A. Geselowitz, T. Greer, B. ter Haar Romeny, J. B. Zimmerman, and K. Zuiderveld. Adaptive histogram equalization and its variations. *Computer Vision, Graphics, and Image Processing*, 39(3):355–368, sep 1987.

[11] E. A. Sickles, C. J. D'Orsi, L. W. Bassett, et al. *ACR BI-RADS® Atlas, Breast Imaging Reporting and Data System*. American College of Radiology, Reston, US, 5th edition, 2013.

[12] D. Tanaka, D. Ikami, T. Yamasaki, and K. Aizawa. Joint Optimization Framework for Learning with Noisy Labels. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5552–5560, jun 2018.

[13] D. Ulyanov, A. Vedaldi, and V. Lempitsky. Instance Normalization: The Missing Ingredient for Fast Stylization. *arXiv pre-prints*, jul 2016.

[14] N. Wu, K. J. Geras, Y. Shen, J. Su, S. G. Kim, E. Kim, S. Wolfson, L. Moy, and K. Cho. Breast Density Classification with Deep Convolutional Neural Networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6682–6686, apr 2018.