A deep learning method for classifying mammographic breast density categories

Aly A. Mohamed
Department of Radiology, University of Pittsburgh School of Medicine, 4200 Fifth Ave, Pittsburgh, PA 15260, USA

Wendie A. Berg
Department of Radiology, University of Pittsburgh School of Medicine, 4200 Fifth Ave, Pittsburgh, PA 15260, USA
Magee-Womens Hospital of University of Pittsburgh Medical Center, 300 Halket St Pittsburgh, PA 15213, USA

Hong Peng
Department of Radiology, Chinese PLA General Hospital, 28 Fuxing Rd Haidian District Beijing 100853, China

Yahong Luo
Department of Radiology, Liaoning Cancer Hospital & Institute, 44 Xiaoheyan Rd Dadong District Shenyang City, Liaoning 110042, China

Rachel C. Jankowitz
Magee-Womens Hospital of University of Pittsburgh Medical Center, 300 Halket St Pittsburgh, PA 15213, USA
Department of Medicine, School of Medicine, University of Pittsburgh, 4200 Fifth Ave, Pittsburgh, PA 15260, USA

Shandong Wu[a)]
Departments of Radiology, Biomedical Informatics, Bioengineering, and Computer Science, University of Pittsburgh, 4200 Fifth Ave, Pittsburgh, PA 15260, USA

(Received 27 June 2017; revised 9 November 2017; accepted for publication 12 November 2017; published 22 December 2017)

Purpose: Mammographic breast density is an established risk marker for breast cancer and is visually assessed by radiologists in routine mammogram image reading, using four qualitative Breast Imaging and Reporting Data System (BI-RADS) breast density categories. It is particularly difficult for radiologists to consistently distinguish the two most common and most variably assigned BI-RADS categories, i.e., "scattered density" and "heterogeneously dense". The aim of this work was to investigate a deep learning-based breast density classifier to consistently distinguish these two categories, aiming at providing a potential computerized tool to assist radiologists in assigning a BI-RADS category in current clinical workflow.

Methods: In this study, we constructed a convolutional neural network (CNN)-based model coupled with a large (i.e., 22,000 images) digital mammogram imaging dataset to evaluate the classification performance between the two aforementioned breast density categories. All images were collected from a cohort of 1,427 women who underwent standard digital mammography screening from 2005 to 2016 at our institution. The truths of the density categories were based on standard clinical assessment made by board-certified breast imaging radiologists. Effects of direct training from scratch solely using digital mammogram images and transfer learning of a pretrained model on a large non-medical imaging dataset were evaluated for the specific task of breast density classification. In order to measure the classification performance, the CNN classifier was also tested on a refined version of the mammogram image dataset by removing some potentially inaccurately labeled images. Receiver operating characteristic (ROC) curves and the area under the curve (AUC) were used to measure the accuracy of the classifier.

Results: The AUC was 0.9421 when the CNN-model was trained from scratch on our own mammogram images, and the accuracy increased gradually along with an increased size of training samples. Using the pretrained model followed by a fine-tuning process with as few as 500 mammogram images led to an AUC of 0.9265. After removing the potentially inaccurately labeled images, AUC was increased to 0.9882 and 0.9857 for without and with the pretrained model, respectively, both significantly higher ($P < 0.001$) than when using the full imaging dataset.

Conclusions: Our study demonstrated high classification accuracies between two difficult to distinguish breast density categories that are routinely assessed by radiologists. We anticipate that our approach will help enhance current clinical assessment of breast density and better support consistent density notification to patients in breast cancer screening. [https://doi.org/10.1002/mp.12683]

Key words: BI-RADS, breast density, convolutional neural network (CNN), deep learning, digital mammography, transfer learning

1. INTRODUCTION

Mammography is the standard screening examination for breast cancer. Breast density is a measure used to describe the proportion of fibroglandular tissue in a woman's breast depicted on a digital mammogram. Breast density can be measured qualitatively or quantitatively. Qualitative methods include the original Wolfe criteria,[1] the Tabar classification,[2] and the widely used Breast Imaging and Reporting Data System (BI-RADS) criteria.[3] The BI-RADS mammographic breast density criteria include four qualitative categories: (a) almost entirely fatty, (b) scattered areas of fibroglandular density (or "scattered density" for short), (c) heterogeneously dense, (d) or extremely dense. Quantitative methods include Cumulus software[4] to interactively determine the skin line and to set a threshold for segmenting dense tissue; the area of dense tissue is expressed in one of six-category percentages: 0, <10%, 10–25%, 26–50%, 51–75%, and >75%.[5–8] Automated computerized methods include the LIBRA program[9] that is publicly available to estimate an area-based percent density as well as the volume-based commercial software, such as Quantra[10] and Volpara.[11] The volume-based methods function only on the raw ("FOR PROCESSING") digital mammogram images, which are not routinely stored in most medical centers.

Several large studies have established that mammographic breast density is an imaging-based risk marker for breast cancer,[12,13] independent of age and menopausal status. Women with extremely dense breasts have a 4–6 fold higher risk compared to women with fatty breasts.[14] When comparing a woman with heterogeneously dense breasts to women with average breast density, her risk is about 1.2 times higher to develop breast cancer; likewise, the risk is about 2.1 times higher when comparing a woman with extremely dense breasts to women with average breast density.[15] Dense breasts are those given a clinical BI-RADS assessment of either "heterogeneously dense" or "extremely dense", and supplemental screening with ultrasound or even magnetic resonance imaging (MRI) may be recommended. Dense breasts indicate a higher risk of masking cancers and reduce the sensitivity. In the U.S., 31 states have enacted breast density notification legislation,[16] and most of those laws require women to receive some level of information regarding their breast density as part of the results letter from their mammograms. Because recommendations for supplemental screening and risk management may vary by breast density, it is highly desirable in the clinic to have a consistent assessment of breast density.

Current assessment of breast density by qualitative BI-RADS categories is subjective, with substantial inter- and intrareader variability;[14,17] it is, however, the standard in current clinical practice. Improving the accuracy and consistency of breast density assessment is an unmet clinical need, whilst there is automated quantitative density assessment emerging and becoming available in limited facilities. While assessment of fatty breasts and extremely dense breasts is highly consistent, there is greater variability distinguishing scattered density from heterogeneously dense parenchyma[17] (Fig. 1). In the 5th edition of BI-RADS, the "heterogeneously dense" assessment may be made based only on one dense area of the breast, emphasizing its potential masking effect. This further increases the variability in the clinical assessment on this BI-RADS density category.

The accuracy of most conventional classification algorithms (e.g., support vector machines) is based on strong feature engineering, which requires prior expert knowledge of the data and a hard hand-crafting process to build descriptive features. Conversely, deep learning can extract features automatically and directly from original data.[18] Deep learning coupled with big training data has shown promising capability in many artificial intelligence applications[19,20] and, more recently, biomedical imaging analysis. For example, deep learning has been used for thoraco-abdominal lymph node detection and interstitial lung disease classification,[21] real-time 2D/3D registration of digitally reconstructed X-ray images,[22] breast lesion detection and diagnosis,[23–27] radiological imaging segmentation,[28,29] as well as digital breast pathology image analysis[27,30] such as mitosis detection and counting, tissue classification (e.g., cancerous vs. non-cancerous), segmentation (e.g., nuclei or epithelium),[30] and metastatic cancer detection from whole slide images of sentinel lymph nodes.[27] Many such studies have shown that automatic feature extraction using deep learning outperformed traditional hand-crafted imaging descriptors.[22,23,31]

Several studies have investigated deep learning in mammographic breast imaging related tasks. Dubrovina et al.[32] presented a novel supervised CNN framework for breast anatomy (i.e., pectoral muscle, dense tissue, and nipple) classification in mammography images, using a patch-wise approach for CNN training. Deep learning has been tested for diagnosis of lesions in several scenarios, such as differentiating between benign and malignant masses;[23] discrimination of masses, microcalcifications, and their combination;[24] between tumor/mass and normal tissue;[26,27] or between three classes of benign, malignant, and normal tissue.[25] A new technique combining a cascade of deep learning and random forest classifiers has been proposed to detect masses in mammogram images.[33] In addition, unlabeled imaging data and unsupervised feature learning (e.g., sparse autoencoder) have been explored for breast density segmentation and risk scoring.[28] Deep convolution and deep belief networks have been integrated in structured prediction models for mammographic breast mass segmentation.[29]

In this paper, we propose a deep learning-based approach using CNN to build a computerized reader for BI-RADS-based breast density categorization. Our work focused on distinguishing between the two most difficult to distinguish BI-RADS density categories, i.e., "scattered density" vs. "heterogeneously dense". Our goals were to evaluate the breast density classification performance of a deep learning-based CNN framework coupled with a large (i.e., 22,000 images) set of digital mammogram images, and to perform a preliminary evaluation on the effects of "transfer learning"[34] in this breast density classification task. Realizing that there

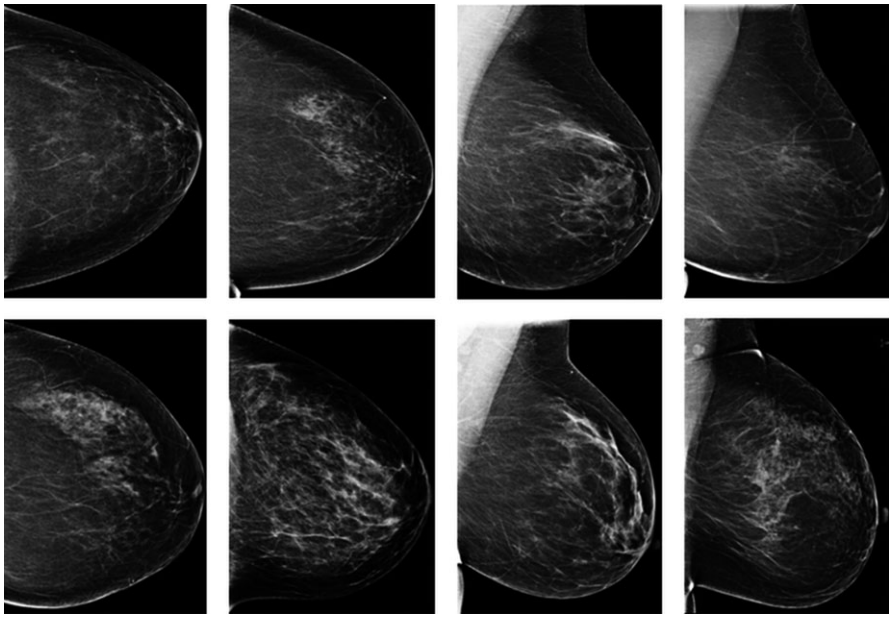

FIG. 1. Examples of digital mammogram images illustrating the two difficult to distinguish Breast Imaging and Reporting Data System (BI-RADS) breast density categories. The top row are breasts with a "scattered density" (Category B) assessment and the bottom row are breasts with a "heterogeneously dense" (Category C) assessment, all made in routine clinical practice.

is essentially no "ground truth" for clinical assessment of breast density, we have leveraged the use of a quantitative assessment of breast density to moderate the robustness of the classification.

2. MATERIALS AND METHODS

2.A. Dataset

We performed a retrospective study that was compliant with the Health Insurance Portability and Accountability Act (HIPAA) and received Institutional Review Board (IRB) approval by the Human Research Protection Office (HRPO) at our institution. Informed consent from patients was waived due to the retrospective nature. From another IBR-approved existing study, we identified a cohort of 1,427 women who underwent standard digital mammography screening from 2005 to 2016 and collected a large dataset of total 22,000 digital mammogram images associated with this cohort. One patient may have multiple (in average 4, range 1–7) sequential screening mammogram examinations (each examination has 4 images). The 22,000 images were those reported to have either a "scattered density" (7,925 images) or "heterogeneously dense" parenchyma (14,075 images) on the clinical mammography report. All the collected mammogram images are negative or breast-cancer free at the time of study. The BI-RADS-based breast density categories that had been routinely assigned in standard clinical workflow by radiologists were retrieved from mammography reports and used as ground truth. The BI-RADS density categories were clinically assessed by a mix of many different board-certified breast imaging radiologists with a varying range of experience in breast imaging. All mammogram examinations were acquired by the Hologic (Marlborough, MA, USA) full-field digital mammography units. Both the mediolateral oblique (MLO) and craniocaudal (CC) views of the left and right breasts on the processed (i.e., "FOR PRESENTATION") images were analyzed.

2.B. Deep learning model

We employed a two-class CNN-based deep learning model, which has shown promising performance in recent work for image classification and pattern recognition.[20] The CNN used an improved version of the AlexNet model, which is not trained with the relighting data-augmentation and the order of the pooling and normalization layers is switched.[35] The CNN structure consists of five convolutional layers, three max-pooling layers and three fully connected layers with a final 2-way softmax function. The two-class CNN model was constructed as an end-to-end system aiming at classifying the two BI-RADS breast density categories: "scattered density" vs. "heterogeneously dense". The CNN model was implemented using the Caffe platform running on a desktop computer system with the following specifications: Intel® Core™ i7-4790 CPU@3.60GHZ with 8 GB RAM and a Titan X Pascal Graphics Processing Unit (GPU). In addition to using the GPU to accelerate training, we also used rectified linear units (ReLU) in place of the traditional tangent function and the sigmoid function as the activation function[20] to further speed up the training. We applied 6-fold cross-validation for the inner-loop CNN model training: dividing all the training images randomly into 6 sets with an equal number of images in each set, each time using five sets for training and leaving one set for validation. The validation set is used to calibrate the accuracy of the training process and prevent overfitting.

In the model training process, the optimization of the hyper parameters was performed using a stochastic gradient descent (SGD) method with batch size of 50. In our configuration (weight decay of 0.001 and a momentum of 0.9), we started with a learning rate of 0.001 and dropped the learning rate by a factor of ten every 2,500 iterations. These parameters were fixed in all the experiments.

Several preprocessing steps are applied as follows:

1. The whole-breast region is first separated automatically[9] from the nonbreast region (i.e., air and chest muscles) in each image and then used as the input of the CNN models for training and testing.
2. Histogram equalization was run to adjust the intensity distribution of all images to the same range.
3. All images were resized to a smaller resolution of 227×227 (original resolution: 3518×2800) to allow a higher computational efficiency.
4. The mean image of training data was generated and subtracted from each input image to ensure that every feature pixel has a zero mean.

2.C. Analysis plans

The main analysis used solely our own collected digital mammograms for training (i.e., training from scratch) and also for testing the CNN model. Secondarily, we evaluated the effects of "transfer learning" on mammographic breast density classification. We initialized our CNN model with the weights of the pretrained AlexNet model learned on a very large existing nonmedical imaging database (i.e., ImageNet[36] – a large dataset consisting of 1.2 million labeled nonmedical images for classifying 1,000 classes), followed by a fine-tuning process of the pretrained model using a separate subset of our own mammogram images. The fine-tuning was conducted on all layers of a neural network, and the optimization process was the same with that for model training from scratch (as described in Section 2.B).

In addition, we performed an additional analysis by removing some noisy or potentially inaccurately labeled images to refine our mammogram image dataset. This was done through calculating a quantitative measure of the breast density and comparing it to the clinically-assigned BI-RADS density categories. Here, the quantitative breast percentage density (PD%) was computed using a fully automated computer tool (i.e., LIBRA[9]) for each image. In general, for images with a "scattered density" (Set 1) or with a "heterogeneously dense" (Set 2) category, by definition of the BI-RADS density assessment, it is generally expected that the quantitative PD% in Set 1 will be statistically lower than the quantitative PD% in Set 2. It should be noted that we were not making this a strict and hard condition because it is not a fully validated premise; but in practice for most cases this condition should make sense. Based on this condition, we removed the "scattered density" images where the PD% was "unexpectedly" greater than the average of the PD% of the "heterogeneously dense" images, and likewise, the "heterogeneously dense" images were removed where the PD% was "unexpectedly" lower than the average of the "scattered density" images. The removed cases may reflect some potentially inaccurately labeled data, and by excluding them, we expect that the remaining data are less affected by noisy or inaccurate qualitative BI-RADS density assessment made by visual observations of radiologists. We then repeated the breast density classification on the remaining dataset and compared the classification accuracy to that using the original full dataset.

2.D. Statistical analysis

We used receiver operating characteristic (ROC) analysis with the area under the ROC curve (AUC)[37] to measure the performance of the classifier. In order to account for the clustering/correlation effects caused by the multiple examinations per patient, N. Obuchowski's test[38] was specifically employed to measure the statistical significance of the differences between AUCs of different CNN models. Also, the Bonferroni correction was applied to adjust the *P*-values for multiple comparisons. In order to reduce the potential bias in training the CNN models caused by substantially unbalanced sample sizes for the two categories, we used 7,000 "scattered density" images and 7,000 "heterogeneously dense" images, all randomly selected from the entire dataset of each category, for training the CNN classifier. A separate unseen set of 1,850 images (including 925 with a "scattered density" and 925 with a "heterogeneously dense" category), randomly selected from the remaining images of each category and without overlap with the selected training images, were used for testing the classification performance. The training and testing were repeated 20 times each time with randomly selected training and testing samples and the averaged CNN classification AUCs were reported. Furthermore, in order to test the influence of different sizes of training samples on the classification accuracy, we re-ran the experiments by decreasing the number of training images gradually from 7,000 to 500, while maintaining the same hyper parameters for each subset of the training samples and using the same number of testing samples each time (i.e., 925 for each category).

3. RESULTS

3.A. Training from scratch on our own mammogram images

We first present the main analysis of our CNN-based breast density classification model, trained directly and solely by using our own mammographic data. Figure 2 shows selected representative ROC curves for two different training sample sizes, while the AUCs with 95% confidence intervals (CIs) with respect to a range of training sample sizes from 500 to 7,000 were plotted in Fig. 3. When trained on 500 images for each of the two classes, the AUC of our CNN model for 1,850 unseen testing samples was 0.9081. In comparison, AUC increased to 0.9421 when trained on 7,000

images. It was observed from Fig. 3 that in general the classification AUC gradually improved along with the increase in the training sample size.

3.B. Effects of transfer learning

Results of using transfer learning are shown in Figs. 4 (representative ROC curves) and 5 (AUCs with 95% CIs when using different number of training samples for fine-tuning). Effects of transfer learning can be seen by comparing to the training from scratch. AUC was improved from 0.9081 (Fig. 3; training from scratch on 500 training samples) to 0.9265 (Fig. 5; when using the pretrained model coupled with 500 mammogram images for fine-tuning). However, when the largest size of training samples was used (i.e., 7,000 images), the AUC (0.9243 in Fig. 5) was slightly lower ($P = 0.166$) than when not using the pretrained model (0.9421 in Fig. 3). Figure 5 shows that the classification performance is relatively stable (around approximately 0.92) when the sample size varied gradually from 500 to 7,000.

3.C. Additional analysis on removing noisy data

According to the criteria defined for additional analysis in Section 2.C, of the 7,925 "scattered density" images there were 867 images (i.e., 10.9%) whose PD% was greater than the average PD% (i.e., 15.0%) of the 14,075 "heterogeneously dense" images; similarly, out of the 14,075 "heterogeneously dense" images, there were 2,286 images (i.e., 16.2%) whose PD% was lower than the average PD% (i.e., 29.5%) of the 7,925 "scattered density" images. We removed the 867 images from the "scattered density" images and 2,286 from the "heterogeneously dense" images, respectively, leaving a total of 18,847 images compared to the original full set of 22,000 images. Then we repeated the classification using the

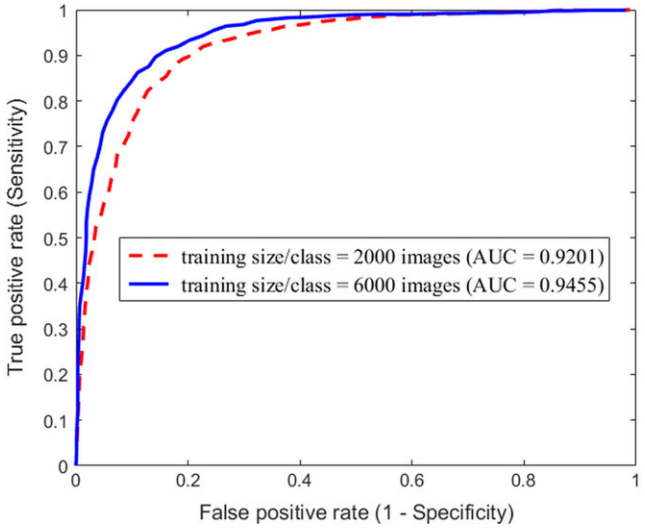

FIG. 2. Representative ROC curves on breast density classification performance for the convolutional neural network (CNN) model trained from scratch on our digital mammogram images. [Color figure can be viewed at wileyonlinelibrary.com]

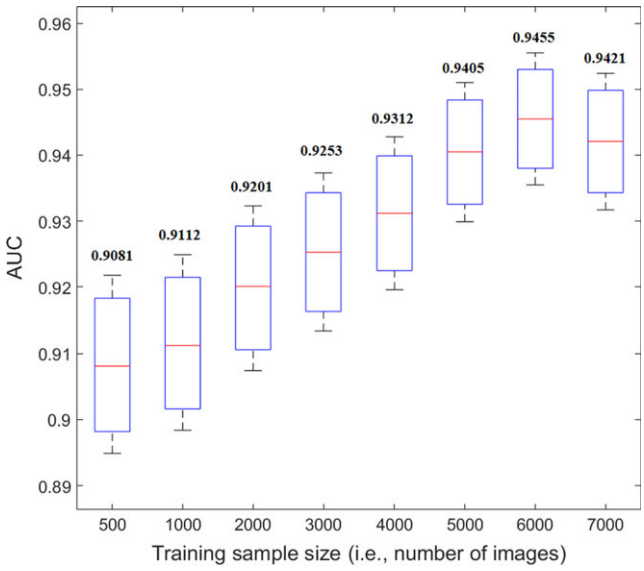

FIG. 3. Boxplot of breast density classification AUCs with 95% CIs for the convolutional neural network (CNN) model trained from scratch on our mammogram images. A general trend of improved AUC with increased training samples was observed. [Color figure can be viewed at wileyonlinelibrary.com]

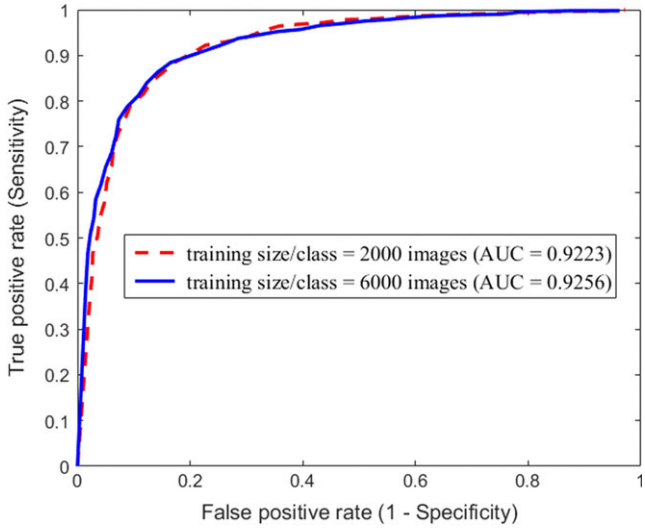

FIG. 4. Representative ROC curves on breast density classification performance when using the pretrained convolutional neural network (CNN) model with a fine-tuning process. [Color figure can be viewed at wileyonlinelibrary.com]

remaining 18,847 images, where 6,000 were randomly selected for each category for training and 1,850 (including 925 "scattered density" and 925 "heterogeneously dense") randomly selected images (no overlap with training samples) for testing. As shown in Fig. 6, after removing those "noisy" images, AUC was increased to 0.9882 and 0.9857 for without and with the pretrained model respectively. Recall that the corresponding AUC was 0.9455 (Fig. 2) and 0.9256 (Fig. 4), respectively, when 6,000 training samples were used before the noisy image removal. The increase in AUC from 0.9455 (Fig. 2) to 0.9882 (Fig. 6) is statistically significant

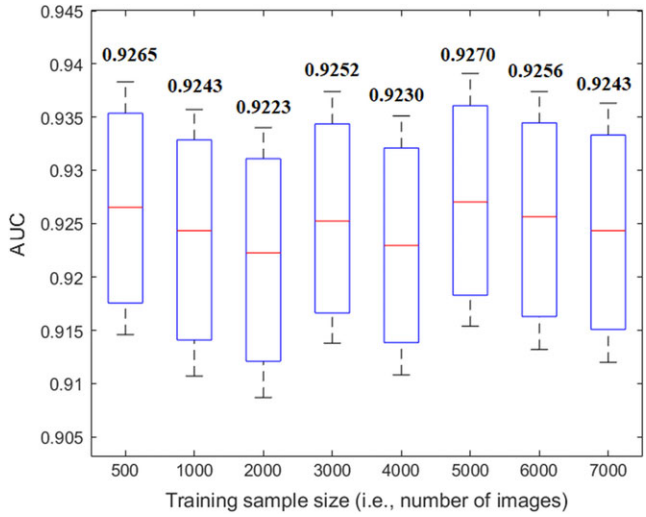

FIG. 5. Boxplot of breast density classification AUCs with 95% CIs when using the pretrained convolutional neural network (CNN) model with a fine-tuning process. A relatively stable classification performance was observed when using a different number of training samples for fine-tuning the models. [Color figure can be viewed at wileyonlinelibrary.com]

($P < 0.001$) without transfer learning. Similarly, the increase in AUC from 0.9256 (Fig. 4) to 0.9857 (Fig. 6) is also statistically significant ($P < 0.001$) with transfer learning. All these comparisons showed that after removing the noisy or potentially inaccurately labeled images, the CNN-based breast density classification performance was significantly improved compared to without the removal.

4. DISCUSSION

In this work, we presented a new deep learning-based medical imaging application to distinguish between the two most difficult to distinguish "scattered density" and "heterogeneously dense" categories in clinical breast density assessment. Assessing breast density is a routine clinical need for a large amount of digital mammogram images acquired in breast cancer screening. Accurate and consistent breast density assessment is highly desirable in the current context of breast density notification in order to provide more informed clinical decision-making support to both clinicians and patients. Out of the four BI-RADS breast density categories, it is fairly easy to distinguish the categories of "almost entirely fatty" and "extremely dense" by visual assessment, where radiologists are realistically comfortable to make decisions without needing assistance. However, it is challenging for radiologists in assigning a consistent classification of "scattered density" vs. "heterogeneously dense", due to the difficulty of discerning the visual features of dense breast tissues between the two categories. Therefore, this work focused on building a CNN model to assist the radiologists in assigning the two difficult to distinguish categories. This represents a more targeted development and use of computerized tools to meet realistic clinical needs. In the potential use of this CNN model integrated in clinical workflow as a second reader, if radiologists are in doubt between the two categories, they could use this tool to generate a prediction to help improve determination of a BI-RADS breast density category.

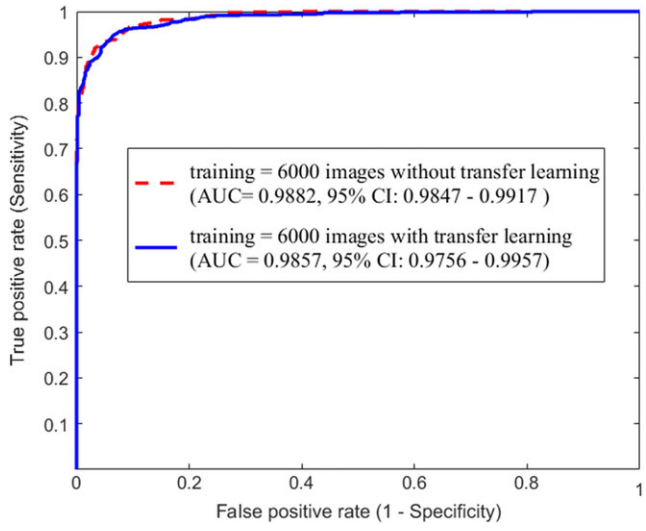

FIG. 6. Breast density classification performance after removing potentially inaccurately labeled images. The AUCs with 95% CIs showed a significant ($P < 0.001$) improvement over using the original full imaging dataset. [Color figure can be viewed at wileyonlinelibrary.com]

We collected a large mammogram imaging dataset and showed that the CNN-based classifier can achieve the highest AUC of 0.94–0.98. Overall, the more training samples were used, the higher the AUC achieved. At the same time, it is observed that a much smaller size of training samples, as small as 500, also led to a just slightly lower accuracy. This may indicate that the CNN-based deep learning approach was able to effectively identify essential imaging features from a relatively small number of training images, to distinguish the two BI-RADS breast density categories.

We evaluated effects of transferring prelearned knowledge on a very large nonmedical image dataset (i.e., ImageNet, >1 million nature images with labeled ground truth) and then fine-tuning the pretrained model with a subset of the specific mammogram images in our breast density-based task. This analysis was motivated by the fact that a large set of reliably labeled medical images (such as outlined ground truth of lesion, tissue, or anatomy) is hard and time-consuming to generate manually by experts and therefore, not commonly available as benchmark for training.[20] In this work, it is observed that the CNN models based on transfer learning can achieve a comparable classification performance to that without using the transfer learning. It is also seen that in our task the AUC seems not to be very sensitive to the size of the fine-tuning samples, indicating that the pretrained model might already be able to dominate the classification effects and requires only minimal fine-tuning optimization. Also, it should be noted that even the fine-tuned performance looks slightly lower than the model trained from scratch, the statistical analysis showed that the difference of the AUCs is not significantly different. So, they are still in comparable range. In addition, in the second experiment, after removing some noisy data, the AUC is still very close to each other with or

without transfer learning. All these results may indicate some unique characteristics of this dataset, the specific task, or the role ImageNet has played in our experiments. But essentially these findings will need further investigations, especially by testing the models on a larger and external dataset from other institutions.

In this work, we used an end-to-end deep learning approach for the density category classification. In testing the transfer learning effects, we used the AlexNet model pretrained on ImageNet and fine-tuned it with our own mammogram data. Another method of transfer learning is to use a deep learning model as a feature extractor. For example, Bram van Ginneken et al.[39] used features extracted from the first fully connected layer of the CNN and then fed these features into a linear support vector machine for classification. Such off-the-shelf features extracted from off-the-shelf deep learning models trained on natural images have shown competitive performance on certain medical tasks.[39] These are interesting findings regarding the usefulness of transfer learning from nonmedical images to medical imaging-based applications, and we believe that further investigation on this topic is still merited.

We realize that there is essentially no ground truth for the BI-RADS-based visual breast density assessment. Radiologists may not be able to consistently reproduce their assessment, and within or amongst different radiologists, there are frequent between "scattered density" and "heterogeneously dense" assignments. Also, clinics demand objective and reproducible assessment of breast density, and there are attempts to use some of the automated computerized algorithms (such as LIBRA, Quantra, and Volpara) to generate quantitative breast density measures. While this is considered the trend of the future, these algorithms either require further evaluation (such as LIRBA) or are limited to specific setting (for example, Quantra and Volpara need to use raw mammogram images). At this stage of the clinical reality, BI-RADS-based breast density categories are still the standard/mainstream for breast density assessment. We emphasize that our study is valuable because it not only used real-world clinical images and the breast density assessment annotated in standard clinical workflow, but also addresses a real clinical question and furthermore, served as another clinical application that showed the capability of the newly emerged revolutionary deep learning techniques. Our CNN model is mainly to be used to assist radiologists in the two BI-RADS category determination, and as clinically needed, it could certainly be used to complement the quantitative breast density assessment software as well. In addition, we performed an additional analysis by removing certain "noisy" images – those that may represent some of the less-reliably or inaccurately labeled cases in actual clinical practice when determining the two BI-RADS density categories (e.g., one category might be accidently or "wrongly" assigned to the other). After excluding those images, the classification performance was boosted to even higher AUCs (around 0.98), with or without the transfer learning. This showed that overall our CNN-based deep learning models were able to automatically and precisely distinguish the two difficult to distinguish BI-RADS breast density categories.

The strengths of our studies include (a) usage of a large clinical imaging dataset, (b) test of training from scratch and transfer learning of pretrained deep learning models, and (c) additional analysis to mitigate the influence of noisy or wrongly labeled data (i.e., the assigned BI-RADS density categories). Our study has some limitations. First, this is a single center retrospective study and our data have little variation in the imaging acquisitions in terms of mammographic vendors and imaging parameters. Second, the studied images were read by many radiologists and we did not track which radiologist interpreted which images (the labor of this task is beyond our affordable efforts). However, if we were to have such information, it would enable us to evaluate the reading performance of different readers and potentially use that insight as a means to enhance understanding to their reading behaviors, and accordingly improve training/education on mammographic image reading. In addition, the lack of ground truth for breast density assessment is a limitation. However, after removing the potentially inaccurately labeled images, among the remaining large number (i.e., 18,847) of images, a higher consistency between the BI-RADS-based truth read by radiologists and the PD% generated by the LIBRA software can be reasonably expected. Finally, this study included only cancer-free mammogram images, considering that use of cancer-affected images may introduce bias in training a breast density classification model (e.g., the model may unexpectedly learn features associated with tumor other than dense tissue). Nevertheless, we plan in our future work to evaluate the effects of the learned model in this study by testing (not training) it using cancer-affected images. In general, our study will benefit from more in-depth analysis and larger scale evaluation. We plan to further improve our study by, for example, identifying which type of images are more likely to be misclassified, comparing other CNNs model structures to the AlexNet, developing other strategies to deal with noisy or potentially inaccurately labeled data, testing by using larger multi-center datasets, and so on.

Deep learning has shown an excellent capability in learning imaging features/traits from annotated imaging data without the need of conventional hand-crafting feature engineering. Feature engineering can be hard especially for the studied breast density assessment problem because it is not straightforward to directly model how radiologists make the visual decision and what kind of local and/or holistic imaging information they rely on in qualitatively determining either of the two BI-RADS breast density categories. This also limits a further comparison between the deep learning-based method and feature engineering-created descriptors. In this study, we demonstrated a deep learning-based CNN model that automatically identifies discriminative information from clinically annotated mammogram images and these images were read by many radiologists in a long-term clinical practice. Our approach avoids hand-crafting processes of image features and therefore is expected to generate more consistent breast density assessment to potentially help improve current clinical procedures in breast density assessment and notification.

5. CONCLUSION

Our study showed encouraging classification performance by a deep learning-based CNN model in distinguishing the "scattered density" vs. "heterogeneously dense" breast density categories. This work adds a new example of applying deep learning and transfer learning in analyzing a large clinical breast imaging dataset. We anticipate that our approach will provide a promising computerized toolkit to help enhance current clinical assessment of breast density and better support lawful density notification to patients in breast cancer screening.

ACKNOWLEDGMENTS

This work was supported by a National Institutes of Health (NIH)/National Cancer Institute (NCI) R01 grant (1R01CA193603), a Radiological Society of North America (RSNA) Research Scholar Grant (RSCH1530), a University of Pittsburgh Cancer Institute Precision Medicine Pilot Award from The Pittsburgh Foundation (MR2014-77613), and a NIH grant (5UL1TR001857) from the National Center for Advancing Translational Sciences (NCATS). We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan X Pascal GPU used for this research.

[a]Author to whom correspondence should be addressed. Electronic mail: wus3@upmc.edu; Telephone: + (412) 641-2567; Fax: + (412) 641-2582.

REFERENCES

1. Wolfe JN. Breast patterns as an index of risk for developing breast cancer. *Am J Roentgenol*. 1976;126:1130–1137.
2. Gram IT, Funkhouser E, Tabár L. The Tabar classification of mammographic parenchymal patterns. *Eur J Radiol*. 1997;24:131–136.
3. Sickles EA, D'Orsi CJ, Bassett LW, et al. *ACR BI-RADS® Atlas, Breast Imaging Reporting and Data System*. Reston, VA: American College of Radiology; 2013:39–48.
4. Byng JW, Boyd NF, Fishell E, et al. The quantitative analysis of mammographic densities. *Phys Med Biol*. 1994;39:1629–1638.
5. Maskarinec G, Pagano I, Lurie G, et al. A longitudinal investigation of mammographic density: the multiethnic cohort. *Cancer Epidemiol Biomarkers Prev*. 2006;15:732–739.
6. Habel LA, Capra AM, Oestreicher N, et al. Mammographic density in a multiethnic cohort. *Menopause*. 2007;14:891–899.
7. Stone J, Dite GS, Gunasekara A, et al. The heritability of mammographically dense and nondense breast tissue. *Cancer Epidemiol Biomarkers Prev*. 2006;15:612–617.
8. Boyd NF, Lockwood GA, Martin LJ, et al. Mammographic densities and breast cancer risk. *Breast Disease*. 1998;10:113–126.
9. Keller BM, Nathan DL, Wang Y, et al. Estimation of breast percent density in raw and processed fullfield digital mammography images via adaptive fuzzy C-means clusteringand support vector machine segmentation. *Med. Phys*. 2012;39:4903–4917.
10. Ciatto S, Bernardi D, Calabrese M, et al. A first evaluation of breast radiological density assessment by QUANTRA software as compared to visual classification. *Breast*. 2012;21:503–506.
11. Youk JH, Gweon HM, Son EJ, et al. Automated Volumetric Breast Density Measurements in the Era of the BI-RADS Fifth Edition: A Comparison With Visual Assessment. *AJR Am J Roentgenol*. 2016;206:1056–1062.
12. Boyd NF, Guo H, Martin LJ, et al. Mammographic density and the risk and detection of breast cancer. *N Engl J Med*. 2007;356:227–236.
13. McCormack VA, Silva IS. Breast density and parenchymal patterns as markers of breast cancer risk: a meta-analysis. *Cancer Epidemiol Biomarkers Prev*. 2006;15:1159–1169.
14. Huo CW, Chew GL, Britt KL, et al. Mammographic density—a review on the current understanding of its association with breast cancer. *Breast Cancer Res Treat*. 2014;144:479–502.
15. Kerlikowske K, Cook AJ, Buist DS, et al. Breast cancer risk by breast density, menopause, and postmenopausal hormone therapy use. *J Clin Oncol*. 2010;28:3830–3837.
16. Mohamed AA, Luo Y, Peng H, et al. Understanding clinical mammographic breast density assessment: a deep learning perspective. *J Digit Imaging*. 2017. https://doi.org/10.1007/s10278-017-0022-2.
17. Berg WA, Campassi C, Langenberg P, et al. Breast imaging reporting and data system: inter-and intraobserver variability in feature analysis and final assessment. *Am J Roentgenol*. 2000;174:1769–1777.
18. Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. *Med Image Anal*. 2017;42:60–48. arXiv preprint arXiv. 1702.05747.
19. Deng L, Yu D. Deep learning: methods and applications. *Foundations Trends Signal Process*. 2014;7:197–387.
20. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521:436–444.
21. Shin H, Roth HR, Gao M, et al. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Trans Med Imaging*. 2016;35:1285–1298.
22. Miao S, Wang ZJ, Liao R. A CNN regression approach for real-time 2D/3D registration. *IEEE Trans Med Imaging*. 2016;35:1352–1363.
23. Cheng J, Ni D, Chou Y, et al. Computer-aided diagnosis with deep learning architecture: applications to breast lesions in us images and pulmonary nodules in CT scans. *Sci Rep*. 2016;6:24454.
24. Wang J, Yang X, Cai H, et al. Discrimination of breast cancer with microcalcifications on mammography by deep learning. *Sci Rep*. 2016;6:27327.
25. Gallego J, Montoya D, Quintero O. Detection and diagnosis of breast tumors using deep convolutional neural networks. Conference Proceedings of the XVII Latin American Conference on Automatic Control; 2016: 11–17.
26. Suzuki S, Zhang X, Homma N, et al. Mass detection using deep convolutional neural network for mammographic computer-aided diagnosis. IEEE SICE; 2016: 1382–1386.
27. Wang D, Khosla A, Gargeya R, et al. Deep learning for identifying metastatic breast cancer. arXiv preprint arXiv:1606.05718, 2016.
28. Kallenberg M, Petersen K, Nielsen M, et al. Unsupervised deep learning applied to breast density segmentation and mammographic risk scoring. *IEEE Trans Med Imaging*. 2016;35:1322–1331.
29. Dhungel N, Carneiro G, Bradley AP. Deep learning and structured prediction for the segmentation of mass in mammograms. In: International Conf. Med. Image Computing and Computer-Assisted Intervention; 2015: 605–612.
30. Janowczyk A, Madabhushi A. Deep learning for digital pathology image analysis: a comprehensive tutorial with selected use cases. *J Pathol Inform*. 2016;7:29.
31. Farabet C, Couprie C, Najman L, et al. Learning hierarchical features for scene labeling. *IEEE Trans PAMI*. 2013;35:1915–1929.
32. Dubrovina A, Kisilev P, Ginsburg B, et al. Computational mammography using deep neural networks. *Comp Methods Biomech Biomed Eng Imaging Vis*. 2016;1–5.
33. Dhungel N, Carneiro G, Bradley AP. Automated mass detection in mammograms using cascaded deep learning and random forests. IEEE DICTA; 2015: 1–8.
34. Zhou Z, Shin J, Zhang L, et al. Fine-tuning convolutional neural networks for biomedical image analysis: actively and incrementally. In: Proc. IEEE CVPR; 2017.
35. Krizhevsky A, Sutskever I, Hinton GE. ImageNet Classification with Deep Convolutional Neural Networks. *Adv Neural Inf Process Syst*. 2012;25:1097–1105.
36. Deng J, Dong W, Socher R, et al. ImageNet: a large-scale hierarchical image database. In: Proc. IEEE Computer Vision and Pattern Recognition; 2009.
37. Fawcett T. An introduction to ROC analysis. *Pattern Recogn Lett*. 2006;27:861–874.
38. Zhou X, McClish DK, Obuchowski NA. *Statistical Methods in Diagnostic Medicine*. Hoboken, NJ: John Wiley & Sons; 2009:569.
39. Ginneken B, Setio AA, Jacobs C, et al. Off-the-shelf convolutional neural network features for pulmonary nodule detection in computed tomography scans. In: Proc. IEEE ISBI; 2015: 286–289.