Multi-View Mammographic Density Classification by Dilated and Attention-Guided Residual Learning

Cheng Li [ID], Jingxu Xu [ID], Qiegen Liu [ID], *Member, IEEE*, Yongjin Zhou [ID], Lisha Mou, Zuhui Pu, Yong Xia [ID], Hairong Zheng [ID], *Senior Member, IEEE*, and Shanshan Wang [ID], *Senior Member, IEEE*

Abstract—Breast density is widely adopted to reflect the likelihood of early breast cancer development. Existing methods of mammographic density classification either require steps of manual operations or achieve only moderate classification accuracy due to the limited model capacity. In this study, we present a radiomics approach based on dilated and attention-guided residual learning for the task of mammographic density classification. The proposed method was instantiated with two datasets, one clinical dataset and one publicly available dataset, and classification accuracies of 88.7 and 70.0 percent were obtained, respectively. Although the classification accuracy of the public dataset was lower than the clinical dataset, which was very likely related to the dataset size, our proposed model still achieved a better performance than the naive residual networks and several recently published deep learning-based approaches. Furthermore, we designed a multi-stream network architecture specifically targeting at analyzing the multi-view mammograms. Utilizing the clinical dataset, we validated that multi-view inputs were beneficial to the breast density classification task with an increase of at least 2.0 percent in accuracy and the different views lead to different model classification capacities. Our method has a great potential to be further developed and applied in computer-aided diagnosis systems. Our code is available at https://github.com/lich0031/Mammographic_Density_Classification.

Index Terms—Radiomics, mammographic density, residual learning, dilated convolution, attention, classification

---

1 INTRODUCTION

Breast cancer is a big heath threat [1], [2], presenting an increasing incidence and a decreasing death rate in all age groups in the past decades [2], [3]. Improved breast cancer treatments and widespread adoption of breast cancer screening techniques, especially mammography, contribute significantly to this favorable trend of mortality reduction [3]. Mammography is the most common and efficient method for breast cancer screening. Clinical studies reported that in addition to mammographic abnormalities (e.g., masses, calcification, architectural distortion, asymmetries), the change of breast density is also an important indicator of early breast cancer development [4], [5], [6].

The very first research on the importance of breast density began with Wolfe *et al.*, who demonstrated the relationship between mammographic parenchymal patterns and the risk of developing breast cancer [7]. Following this, Boyd *et al.* showed a similar correlation between mammographic density and breast cancer risks [8]. Inspired by these discoveries, a number of studies on breast density classification emerged. The American College of Radiology (ACR) Breast Imaging Reporting and Data System (BI-RADS) groups breasts into four categories according to the density with BI-RADS I refers to the lowest densities and BI-RADS IV refers to the highest (BI-RADS I: fat breast (0-25 percent), BI-RADS II: fat with some fibroglandular tissue (26-50 percent), BI-RADS III: heterogeneously dense breast (51-75 percent), and BI-RADS IV: extremely dense breast (76-100 percent)). Patients with extremely dense breasts (BI-RADS IV) have a 2-6 times higher risk of developing breast cancer than patients with fatty breasts (BI-RADS I) [9]. Therefore, breast density plays an important role in the early detection of breast cancer. However, inspection of the generated large quantities of mammograms by radiologists is tedious and subjective, which suffers from poor intra- and inter-radiologists reproducibility [10], [11]. An automatic system which can accurately classify mammographic densities is in urgent need.

- C. Li, H. Zheng, and S. Wang are with Paul C. Lauterbur Research Center for Biomedical Imaging, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China.
  E-mail: {cheng.li6, hr.zheng}@siat.ac.cn, sophiasswang@hotmail.com.
- J. Xu is with Paul C. Lauterbur Research Center for Biomedical Imaging, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China, and also with the School of Biomedical Engineering, Health Science Center, Shenzhen University, Shenzhen 518060, China. E-mail: xujingxu2016@email.szu.edu.cn.
- Q. Liu is with the Department of Electronic Information Engineering, Nanchang University, Nanchang 330047, China.
  E-mail: liuqiegen@hotmail.com.
- Y. Zhou is with the School of Biomedical Engineering, Health Science Center, Shenzhen University, Shenzhen 518060, China.
  E-mail: yjzhou@szu.edu.cn.
- L. Mou and Z. Pu are with the Second People's Hospital of Shenzhen, the First Affiliated Hospital of Shenzhen University, Shenzhen 518035, China. E-mail: lishamou@gmail.com, pupeter190@163.com.
- Y. Xia is with the Shaanxi Key Lab of Speech and Image Information Processing (SAIIP) and the Centre for Multidisciplinary Convergence Computing (CMCC), School of Computer Science and Engineering, Northwestern Polytechnical University, Xi'an 710072, China. E-mail: yxia@nwpu.edu.cn.

Many studies measured breast density by quantifying the gray-level histograms of mammograms [12], [13], [14]. Subsequent studies found that it might be insufficient to classify breasts into the corresponding BI-RADS categories based only on the histogram information. For example, the study by Oliver *et al.* illustrated that the four different categories are quite similar with regard to both the mean gray-level values and the shapes of the histogram [4]. To address this issue, researchers turned to applying traditional feature engineering methods to handle the breast density classification task. A general procedure of feature engineering approaches involves regions of interest (ROIs) segmentation, feature extraction, and classification. Bovis *et al.* got an accuracy of 71.4 percent by using a classifier paradigm where a combination of the Fourier and discrete wavelet transforms was investigated on the first and second-order statistical features [15]. Oliver *et al.* extracted morphological and texture features from breast tissue regions which were segmented using a fuzzy C-means clustering technique, and these features were then treated as inputs for the breast density classifier [4]. Jensen *et al.* adopted the same breast tissue segmentation method but extracted the first and second-order statistical features as well as morphological features for the Mammographic Image Analysis Society (MIAS) dataset [16]. These two studies achieved 86.0 and 91.4 percent breast density classification accuracies, respectively. Chen *et al.* evaluated different local features using texture representation algorithms. After that, they modelled mammographic tissue patterns based on the local tissue appearances in mammograms [17]. The work of Indrajeet *et al.* was based on ROIs manually extracted from images. Then, multi-resolution texture descriptors were extracted from 16 sub-band images which were obtained from second level decomposition through wavelet packet transform [18]. Although satisfactory classification performance could be achieved for the target dataset with these traditional feature engineering methods, prior expert knowledge of the data and a hand-crafting process are necessary, which hinders the general wide adoption of the developed algorithms.

On the other hand, the fast development of the deep learning field offers a promising alternative method for medical image analysis [19], [20], [21], [22], [23]. Convolutional Neural Networks (CNNs) are one type of these deep networks that have already shown excellent performance in image classification, detection, and segmentation. CNNs can learn highly nonlinear relationships between the inputs and outputs without human intervention. Radiomics is an emerging method in recent years that works by extracting large amounts of advanced quantitative features from medical images and quantifying the predictive or prognosis relationships between images and medical outcomes according to the features [24], [25]. Nevertheless, the advantages of CNNs have not been fully integrated with the radiomics approach to solve the problems encountered during classifying mammographic density into the four BI-RADS categories. Therefore, in this paper, we propose a CNN-based radiomics method for the automatic extraction of high-throughput features from mammographs and the subsequent classification of the breast densities. Specifically, our contributions can be summarized as follows: (1) We collected a clinical mammographic dataset from 500 patients for the evaluation of breast density classification models. (2)

We proposed an enhanced mammographic density classification model with dilated and attention-guided residual learning. (3) Targeting at analyzing the multi-view mammograms, we specifically designed an effective multi-stream network architecture. (4) Evaluating the multi-view input network architecture on the clinical dataset, we identified that the two imaging views (craniocaudal (CC) view and mediolateral oblique (MLO) view) lead to different classification capabilities of the model. (5) We empirically proved that sharing parameters among the different streams for feature extraction are more beneficial to the classification accuracy. (6) Our code is made publicly available such that our method can serve as a baseline for the following research on mammographic breast density classification.

The rest of this paper is organized as follows. Section 2 lists the related works. Section 3 describes the dataset and the network architecture, as well as the implementation details. In Section 4, the results are presented. Sections 5 and 6 are the discussion and conclusion.

2 RELATED WORKS

A number of studies have applied CNNs to mammogram-related tasks, including breast lesion detection, benign and malignant breast mass differentiation, microcalcification recognition, and their combinations [26], [27], [28], [29], [30], [31]. For the specific topic of mammographic density classification, CNNs have also made great contributions. Fonseca *et al.* utilized CNNs to extract features from mammograms and conducted density classification using a support vector machine classifier [32]. Ahn *et al.* designed a CNN architecture to classify the input mammogram patches into dense or fatty tissues with the help of three additional created images, and estimate the mammographic density by combining the results of all patches [33]. Li *et al.* developed a similar method to classify mammographic density based on a sliding window segmentation approach [34]. These two studies require the manual segmentation maps of dense regions as references. Wu *et al.* treated concatenated pixel intensity histograms of four views (left CC, right CC, left MLO, and right MLO) as the image features and use a three-layer CNN (one input layer, one fully connected layer, and one softmax classification layer) to predict the density categories [35]. Mohamed *et al.* designed an eight-layer CNN (improved from AlexNet) to directly classify the mammograms into two categories (scattered density and heterogeneously dense) [36]. Lehman *et al.* adopted ResNet18 to do mammographic density classification and for the first time, they implemented the deep learning-based model in the clinical application and tested the acceptance of its assessment by the radiologists [37]. Furthermore, Kyono *et al.* developed a multi-task system that could predict factors related to breast cancer, including density, conspicuity, etc. and at the same time pick out those cases that the system could not confidently diagnose for further investigation [38]. Gandomkar *et al.* applied inception-V3 network architecture to process mammograms generated by mammography units manufactured by different manufacturers [39]. All these deep CNN-based studies selected a single CNN architecture to perform the classification task
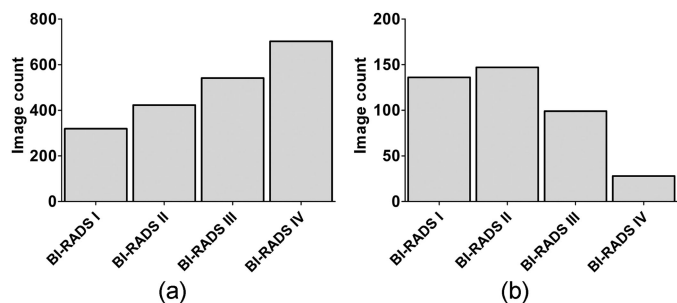
Fig. 1. Image distributions over the four categories. (a) The clinical dataset. (b) The INbreast dataset.

without considering the different capacities of different network structures, which we will investigate in this study.

It is widely recognized that to increase the performance of CNN models, a large receptive field is necessary. To achieve a large receptive field, serial pooling operations or strided convolutions are needed, when severe image resolution decrease happens. This issue is especially troublesome for medical image diagnosis as lesions often occupy small regions of the images. Different methods have been developed to solve this problem, including skip connections to introduce more information from low level feature maps to the final outputs and dilated convolutions. Dilated convolution can preserve the image resolution while increasing the receptive field. Multi-scale contextual information were extracted through utilizing dilated convolutions with different dilation rates to improve the segmentation performance [40], [41]. Dilated convolution was also effective for the image classification task [42]. Another frequently investigated problem with classic CNN models recently is the information redundancy in extracted large number of feature maps. Different attention mechanisms have been proposed to highlight useful information and suppress the redundant one. Attention has been widely used in the natural language processing field. Interestingly, studies found that classic CNNs could implicitly learn to focus on the most important regions of input images [43], which is a kind of spatial attention. Both spatial and channel-wise attention modules have been designed to improve the classification performance of CNNs [44], [45]. Even though many studies have applied deep CNNs to the mammographic density classification task as discussed, the fast updating deep learning methods have not been fully investigated in these aspects. Therefore, we feel motivated to study the feasibility of applying deep residual learning with dilated convolutions and attention mechanisms to mammographic density classification. In addition, it is also interesting to investigate the complementary relationships between the different imaging views of the same patients.

3 MATERIALS AND METHODS

3.1 Dataset

Two datasets were evaluated in this study. First, we conducted a retrospective study on 1985 mammograms collected from 500 female patients. The examinations were performed from October 2015 to August 2018. The patients were aged from 17 to 75. There are 2 to 4 mammograms taken at different views (CC and MLO views) for each patient.
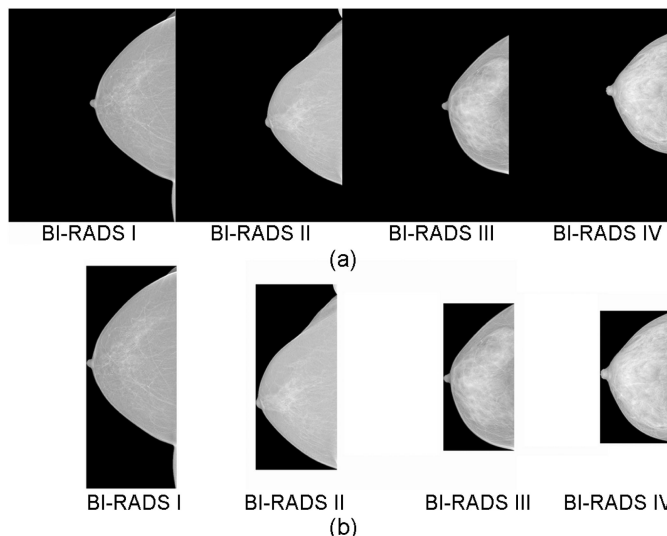

Fig. 2. Example mammograms of the four BI-RADS categories. (a) Mammograms before cropping. (b) Corresponding mammograms after cropping.

Among the 1985 mammograms, 319 were classified as BI-RADS I (86 patients), 423 were BI-RADS II (106 patients), 541 were BI-RADS III (133 patients), and 702 were BI-RADS IV (175 patients) (Fig. 1a). Example images of the four categories are shown in Fig. 2a. This study was approved by the local ethics committee. Informed consent was obtained by all of the involved patients. All examinations and the density classification were performed by radiologists that have decades of experience. One overall BI-RADS category is given to the patient without distinguishing the left or right breast. The mammograms were collected and saved by the standard Digital Imaging and Communications in Medical (DICOM) format. The image matrix has different sizes from 407×1374 to 2394×3062.

Then, we tested our proposed approach on a publicly available dataset, the INbreast dataset [46]. The INbreast dataset was collected from 115 patients (410 images), from which 90 are with both breasts affected (4 images per case) and 25 are mastectomy patients (2 images per case). Two views for each breast were recorded (CC and MLO views). Among the 410 images, 136 were classified as BI-RADS I, 147 as BI-RADS II, 99 as BI-RADS III and 28 as BI-RADS IV (Fig. 1b). The mammograms were acquired on x-ray films and saved by DICOM format. The image matrix has either 3328×4084 or 2560×3328 pixels.

3.2 Data Preprocessing

As introduced in the dataset section, for our clinical dataset, we have a total number of 1985 images. It could be observed from Fig. 2a that the original mammograms contain large regions of irrelevant background. To make full use of the computation power, we cropped out the background regions first (Fig. 2b). Then, the images were resized to 224×224. It is to be noted that our focus is the density classification and the aspect ratios of the mammograms were not very important in this case. Training deep CNN models requires a large number of data and data augmentation is frequently applied. We augmented our training dataset by random horizontal and vertical flippings and random rotations in $[-90^\circ, 90^\circ]$. A 5-fold
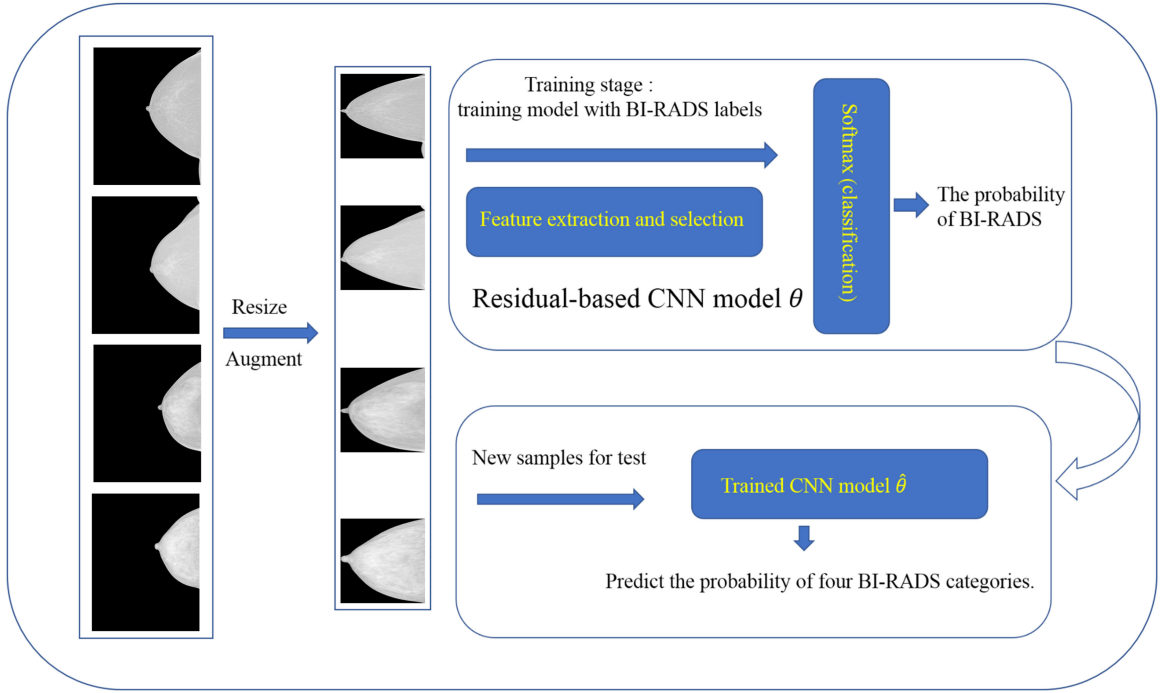
Fig. 3. Schematic diagram of residual learning for classification.

cross-validation experiment was conducted with the dataset. To avoid possible bias, it was guaranteed that all mammograms of the same patient went into either the training dataset or the validation dataset. For the experiments regarding the multi-view inputs, we directly copied the relevant views to manually generate the four inputs if the corresponding views are missing, which leads to a total number of 2,000 images for these experiments.

The same processing steps were applied to the INbreast dataset.

3.3 Deep Residual CNNs

CNNs are a class of deep learning methods that attempt to learn high-level features and tackle the computer vision problems, such as classification, detection, and segmentation. Gradient vanishing is a big problem for deep CNNs. Thanks to the invention of the residual network, CNNs can go substantially deeper now than previous. A detailed description of the residual neural networks could be found in the literature [47]. Briefly, we let the stacked nonlinear layers fit another mapping of the residual $F(X)$ instead of the desired output feature map $H(X)$

$$F(X) = H(X) - X, \qquad (1)$$

and H(X) is recast to

$$F(X) + X. \qquad (2)$$

The formulation of $F(X) + X$ can be realized by feedforward CNN with shortcut connections or skip connections. Due to the propagation of gradients through the shortcut connections, it is easier to optimize the residual mapping of $F(X)$ than to optimize the original mapping of $H(X)$. Therefore, by adding residual learning block, deeper networks

could be designed to extract richer information from images for our classification tasks.

Next, we will describe in detail the CNN method used for image classification. After preprocessing, the training and validation images went through the training and the test stages respectively as shown in Fig. 3. CNNs are trained by feedforward and backpropagation processes. The feedforward process extracts and selects the features and calculates the loss, whereas the backpropagation process optimizes the network parameters by gradient descent of the loss function.

The feedforward process of CNNs could be interpreted by the following steps. First, the images pass through the convolution layers

$$C_l = \sigma_l(W_l * C_{l-1} + b_l), \qquad (3)$$

where l denotes the layer number, σ_l denotes the nonlinear activation (the rectified linear activiation (ReLU) was used for this study), W_l and b_l are weights and bias, $*$ denotes the convolution operation, and C_l denotes the feature maps with C_0 denotes the input. Some convolution layers are followed by downsampling procedure (convolution with strides).

After passing through all the convolution layers, the feature maps were compressed by global average pooling and went through several fully connected (FC) layers. For our classification task, a softmax activation was included after the FC layers

$$C_L = softmax(W_L C_L + b_L), \qquad (4)$$

where $softmax(x)_i = \frac{e^{x_i}}{\sum e^{x_i}}$ and C_L is the output of last FC layer, So the final prediction from the network could be summarized as
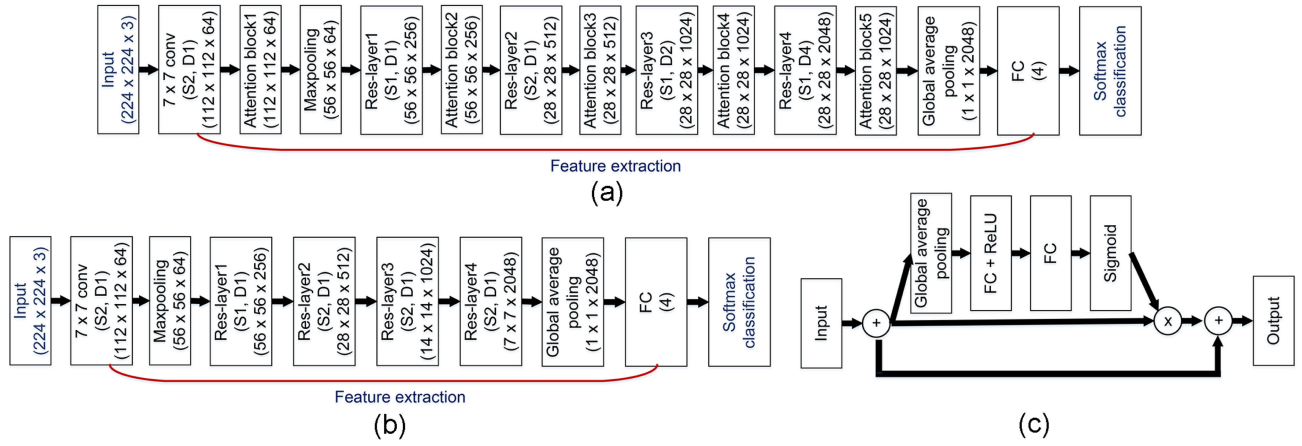
$$Y = C(\theta, X), \qquad (5)$$

Fig. 4. Network architecture. (a) Architecture of the proposed ResNet50 + DC + CA (DC: dilated convolutions. CA: channel-wise attention). (b) Architecture of ResNet50. (c) The attention block. (S1, D1) means that the first convolution in the block has a stride of 1 and all convolutions in the block has a dilation rate of 1. FC refers to fully connected layers.

where θ represents all the network parameters to be optimized, C denotes the overall forward pass network and X refers to the input.

On the other hand, the CNN backward process is the backward propagation of loss gradients, which tries to optimize the network parameters by addressing the following cross-entropy loss minimization problem

$$\hat{\theta} = \arg\min_{\theta} \left\{ - \sum_{k=1}^{K} \left[\sum_{i=1}^{I} Y_i' log(C(\theta, X)) \right] \right\}, \quad (6)$$

where I and K are the total number of classification categories and training samples respectively. Y_i' is the manually labelled ground truth.

After the training phase, a classification model is obtained with the trained parameters. For new independent samples, we can generate the probability distributions of each case by calculating

$$Y_{test} = C(\hat{\theta}, X_{test}). \quad (7)$$

Then the BI-RADS categories of the mammograph images could be determined accordingly.

3.4 Network Architecture With Dilated Convolution and Attention

Our overall network architecture is shown in Fig. 4, which is composed of the input, feature extration, and classification modules. By utilizing dilated convolutions, the resolution of the feature maps before global average pooling is changed to 1/8 of the input resolution (Fig. 4a) instead of 1/32 (Fig. 4b) [47]. A channel-wise attention block was applied to highlight the important feature maps in different layers (Fig. 4c).

Dilated convolution or atrous convolution increases the receptive field of the network without losing the image resolution, which is very important for the medical image diagnosis task. For a dilated convolution with a dilation rate of d, the operation can be represented as

$$(C_l * k)(p) = \sum_{s+d \cdot t = p} C_l(s) k(t), \quad (8)$$

where C_l refers to the feature map at layer l, and k refers to the filter with a dilation rate of d. The receptive field of the element in the output feature map (C_{l+1}) with regards to the input feature map (C_l) is $ks + (ks + 1)(d - 1)$ with ks representing the kernel size. Normal convolution is a special case of dilated convolution with a dilation rate of 1, and increasing the dilation rate enlarges the receptive field linearly.

Each layer in the deep neural networks extracted a large number of features and some of them may be irrelevant to the task. With these irrelevant features sharing the same importance, the important features may be overlooked. The channel-wise attention block in each layer adaptively learns a weight for each feature map in that layer to selectively utilize the extracted feature information. The output of the attention block is

$$Y_c = X_c + S \cdot X_c, \quad (9)$$

where X_c is the input to the attention block and Y_c is the output of the block. S is the weight of each feature map calculated by

$$S = \sigma(W_2 * \delta(W_1 * Z + b_1) + b_2) \quad (10)$$

$$Z = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} (X_c), \quad (11)$$

where σ refers to the Sigmoid function. $W_1 \in R^{n/r \times n}$, $W_2 \in R^{n \times n/r}$, $b_1 \in R^{n/r}$, and $b_2 \in R^n$ are the weights and bias of the FC layers, respectively. $r = 16$ is a reduction ratio.

3.5 Experiments With Multi-View Inputs

To investigate the complementary relationships between the different imaging views, we conducted experiments with multi-view inputs (Fig. 5). In brief, we have performed experiments treating all four views (Fig. 5a) as well as two CC views (Fig. 5b) or two MLO views of the same patients as the inputs. Multiple network streams are utilized to extract features from the different view inputs and the extracted features are concatenated before classification. The different network streams in our experiments can have
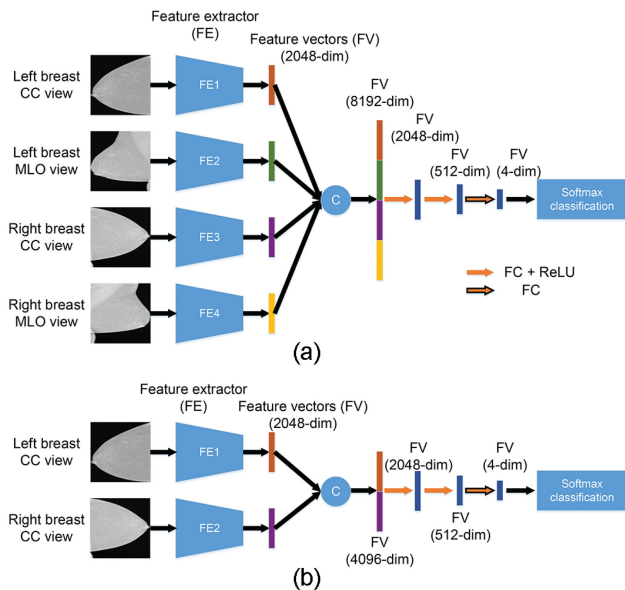
Fig. 5. Network architecture with multi-view inputs. (a) Four-view inputs. (b) Two-view inputs (similar for two MLO-view input setting).

shared parameters or independent parameters, which we will discuss in the results section.

3.6 Parameter Settings and Implementation Details

We used Pytorch to implement our CNN networks for the breast density classification task. Two sets of models were independently trained from scratch for the two datasets. The network training was implemented on a Dell-7910 workstation equipped with two E5-2640v4 Intel Haswells, a NVIDIA TITAN XP GPU and 64G memory. Adam was used for training, with a batch size of 16. The learning rate was initially set to 0.001 and reduced by $\frac{1}{2}$ every 30 epochs. Random values drew from the uniform distribution were used for the weight initialization and zero for the bias initialization. The models with the smallest loss values within 100 training epochs (200 training epochs for the multi-view inputs as slower convergence was observed) were selected as the final models to generate the classification results. We use accuracy, F1 score, and area under the curve (AUC) of receiver operating characteristics curves (ROC) to characterize the classification results. F1 score and AUC are calculated when treating each BI-RADS category classification as a binary classification task, and the final macro-averaged results for all of the four categories are reported.

4 RESULTS

4.1 Results on the Clinical Dataset

4.1.1 Classification Performance With Different Base Residual Networks

First of all, we evaluated the influence of network depth on the classification performance. Table 1 summarizes the classification results of the three different network configurations. It could be found that overall, all the networks showed very similar results with a mild increase in the evaluation metrics of ResNet50. This is acceptable as deeper networks require larger datasets to obtain a proper training despite their larger capacities. For our clinical dataset, ResNet50

TABLE 1
Classification Results of Different Baseline ResNet Models on the Clinical Dataset

Models	Accuracy (%)	F1 score (%)	AUC (%)
ResNet34	86.2	84.7	97.1
ResNet50	**87.1**	**85.6**	**97.2**
ResNet101	86.2	84.7	97.1

might be the one that achieves the best balance between model capacity and dataset requirement. Detailed inspection reveals that for all the three networks, the best classification performance was achieved for BI-RADS IV followed by BI-RADS I, and the worst for BI-RADS II (Table S1 in supplementary file, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TCBB.2020.2970713). This performance difference among the four categories could be explained from two perspectives. On one hand, BI-RADS I and IV are the two most distinguishable categories, so they should be easier to classify. On the other hand, more training images were available for BI-RADS III and IV. Combining these two factors, BI-RADS IV mammograms are the easiest to be accurately classified, whereas BI-RADS II mammograms are the hardest.

To have an in-depth investigation of the classification behavior of the networks, we present the confusion matrix generated from one experiment of ResNet50 (Table 2). It could be observed that except for BI-RADS III, for which several images were misclassified as BI-RADS I (3 images), all the mistakes were made between neighboring categories. For the worst performer BI-RADS II, there were 9 images classified as BI-RADS I, 9 as BI-RADS III, and 0 as BI-RADS IV by the network. As mammographic density classification is rather qualitative and subjective, these mistakes could be tolerated for clinical applications.

4.1.2 Performance Enhancement by Dilated Convolutions and Attention

We tested the effectiveness of the added modules using the best baseline residual networks, ResNet50 (Table 3). Large improvements on accuracy and F1 score were observed whereas AUC results were not influenced. By using dilated convolutions alone, both accuracy and F1 score were increased by 1.1 percent. By adding the channel-wise attention module alone, increases of 1.0 percent on accuracy and of 1.1 percent on F1 score were observed. By simultaneously integrating dilated convolutions and channel-wise attention blocks, accuracy and F1 score were increased by 1.6 and 1.5 percent, respectively, compared to the naive ResNet50.

TABLE 2
Confusion Matrix Generated From One Experiment of ResNet50 on the Clinical Dataset (Accuracy in %)

BI-RADS	I	II	III	IV
I	96.9	3.1	0	0
II	10.1	79.8	10.1	0
III	2.7	7.3	84.5	5.5
IV	0	0	2.1	97.9

TABLE 3
Classification Results of Different Baseline
ResNet Models on the Clinical Dataset

ResNet50	Accuracy (%)	F1 score (%)	AUC (%)
Add DC	88.2	86.7	97.3
Add CA	88.1	86.7	97.2
Add DC & CA	**88.7**	**87.1**	**97.4**

DC: Dilated Convolutions. CA: Channel-wise Attention.

The performance improvement was not a simple summation of the improvement by single modules. For the current network settings, we simply added the attention module after each layer without architecture optimization. As neural architecture search (NAS) is frequently studied recently [48], we plan to optimize our network architecture through NAS in the future.

4.1.3 Comparison to Existing Models

Our proposed ResNet50 with dilated convolutions and attention modules achieved the best classification results when compared to three existing models [36], [37], [39] (Table 4). The detailed results regarding each of the four categories also confirmed that our model could outperform these comparison models (Table S2 in supplementary file, available online).

4.1.4 Effectiveness of Multi-View Inputs

Compare the results utilizing four inputs (Table 5) to those utilizing single-view inputs (Table 3), it can be clearly observed that the four-input setting is beneficial to the classification task and improved statistics were achieved by all the investigated models including the baseline ResNet50. Meanwhile, our proposed network with dilated convolutions and attention modules obtains the highest scores among the four models reflected by all the three evaluation metrics.

For the experiments with two inputs, results indicate that with our clinical dataset under our experimental conditions, networks trained with the two CC-view inputs can perform similar to the networks trained with four inputs, whereas networks trained with the two MLO-view inputs perform much worse even than the single-input models. Therefore, we suggest that CC view is more important for our density classification task. Nonetheless, the proposed network continues to be the best among the four models.

All the above experiments were conducted by sharing the parameters among the different feature extraction streams. When the different streams have independent parameters to optimize, the classification performance was

TABLE 4
Results Comparison to Existing Models on the Clinical Dataset

Models	Accuracy (%)	F1 score (%)	AUC (%)
Mohamed *et al.* [36]	86.2	84.7	97.1
Bahl *et al.* [37]	87.1	85.6	97.2
Gandomkar *et al.* [39]	86.2	84.7	97.1
Ours	**88.7**	**87.1**	**97.4**

TABLE 5
Classification Results of Different Models With
Different Input Settings on the Clinical Dataset

Inputs	ResNet50	Accuracy (%)	F1 score (%)	AUC (%)
Four inputs (CC & MLO)	Baseline	90.3	88.5	97.7
	Add DC	91.1	90.1	97.9
	Add CA	90.9	89.5	97.5
	Add DC & CA	**92.1**	**91.2**	**98.1**
Two inputs (CC)	Baseline	90.3	88.3	97.4
	Add DC	91.1	89.2	97.9
	Add CA	91.1	90.2	97.6
	Add DC & CA	**92.5**	**91.7**	**98.2**
Two inputs (MLO)	Baseline	68.2	61.2	90.2
	Add DC	71.2	66.1	92.1
	Add CA	71.0	64.6	92.8
	Add DC & CA	**75.2**	**67.9**	**93.6**

much worse (Table 6). It might suggest that for our breast density classification task, although different view mammograms contain useful complementary information, the feature extractor should be shared to regularized the feature extraction process.

4.2 Results on the INbreast Dataset

The classification results of the different ResNet-based networks on the INbreast dataset are shown in Fig. 6 (more results in Table S3 in supplementary file, available online). Overall, all the networks showed worse performance on the INbreast dataset than on the clinical dataset but with a similar trend. The results of the three residual networks (ResNet34, ResNet50, and ResNet101) showed no difference in classification accuracy. ResNet34 has slightly higher F1 score and AUC. Dilated convolutions contributed to 4.4 percent increase in classification accuracy but harmed the F1 score and the AUC. Attention modules contributed to 4.9 percent accuracy increase, 8.1 percent F1 score increase, and 1.1 percent AUC increase. Integrating both the dilated convolutions and the attention modules increase the classification accuracy by 7.1 percent, F1 score by 14.1 percent, and AUC by 1.8 percent compared to ResNet50.

Compared to the three existing models, our proposed ResNet with dilated convolutions and attention modules showed the best performance reflected by all the three evaluation metrics (Table 7). Specifically, our model improved the classification accuracy by 6.1 percent, F1 score by 10.4 percent,

TABLE 6
Classification Results of Different Models With Independent
Feature Extraction Streams

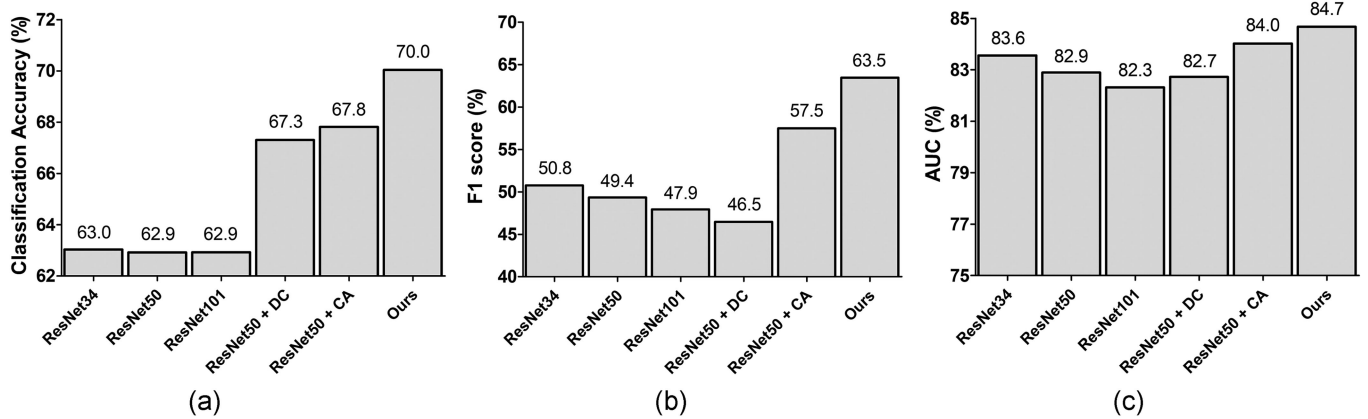| Inputs	ResNet50	Accuracy (%)	F1 score (%)	AUC (%)
Four inputs (CC & MLO)	Baseline	88.7	87.4	97.1
	Add DC & CA	90.3	89.2	97.0
Two inputs (CC)	Baseline	88.9	86.7	97.2
	Add DC & CA	90.3	90.8	97.7
Two inputs (MLO)	Baseline	66.8	62.0	89.0
	Add DC & CA	71.6	63.7	91.4

Fig. 6. Classification results of different networks on the INbreast dataset. (a) Classification accuracy. (b) F1 score. (c) AUC of the ROC.

and AUC by 2.6 percent. Detailed results of each BI-RADS category suggested that under most cases, our proposed model could outperform the existing models (Table S4 in supplementary file, available online).

Similarly, we checked the confusion matrix generated from one experiment of the proposed network on the INbreast dataset (Table 8). Same as the clinical dataset, the most confusing images come from the neighboring categories.

5 DISCUSSION

Traditional radiomics methods extract features based on manual observation and operations, including manually designing the strategies for feature extraction and selection. Compared to the traditional feature engineering approach, deep convolutional networks with residual learning can automatically extract high-order, high-abstraction, and subtle features from mammograms that might even be overlooked by human naked eyes, which enables accurate discrimination of the four BI-RADS categories. Moreover, by working with the whole original images, the classification model has access to all the image-relevant information and elevated performance could be expected. With the proposed single-input dilated and attention-guided deep residual learning network, an overall accuracy of 88.7 and 70.0 percent for the four BI-RADS category classification task were obtained for the clinical dataset and the public dataset, respectively.

The performance of deep neural networks depends significantly on the studied dataset. In our case, we have a larger clinical dataset with relatively balanced distribution of the patients over the four BI-RADS categories. Thus, satisfactory classification results were achieved with all the different ResNet-based models and only a slight improvement was observed for the proposed method. However, for the INbreast dataset, not only the available images were limited, but also a severe imbalanced distribution over the four categories was

observed (Fig. 1b). Training a classification network with this dataset is very challenging. Thus, our proposed model with an increased model capacity achieved significant classification enhancement over the baseline ResNet models. We have tried to balance the category distribution by giving weighted loss during training or augmentation more on BI-RADS IV intensionally. We have also tested to combine the two datasets and train the models on the mixed dataset. In addition, extra experiments with models pretraining on the clinical dataset and finetuning on the INbreast dataset have been conducted. But all these attempts showed no significant performance improvement even though finetuning speeds up the convergence of the model training (Tables S5-S10 in supplementary file, available online). It is acceptable as the two datasets came from different sources and a large domain gap is expected. There are studies showing that utilizing simple weighted loss or joint training for images from different domains lead to decreased model performance [49]. We will go further in this direction by developing strategies to bridge the two datasets or to select those images that can enhance the classification performance of the models when using joint training. It is to be noted that the INbreast dataset contains 107 images with breast masses, which might also affect the density classification. We plan to solve these issues in the following study.

Multi-view mammograms contain complementary information that can be utilized to enhance breast density classification performance. Correspondingly, we designed a multi-stream network architecture and improved classification performance was achieved. Special attention should be paid when designing the multi-input models as our experiments indicate that the CC-view and MLO-view images are not equally beneficial to the density classification task. Our models with two CC-view inputs generate similar or even slightly better results than the four-input models on the clinical dataset. This interesting observation reflects that either

TABLE 7
Results Comparison to Existing Models on the INbreast Dataset

Models	Accuracy (%)	F1 score (%)	AUC (%)
Mohamed *et al.* [36]	59.6	35.4	82.0
Bahl *et al.* [37]	63.8	48.9	81.2
Gandomkar *et al.* [39]	63.9	53.1	82.1
Ours	**70.0**	**63.5**	**84.7**

TABLE 8
Confusion Matrix Generated From One Experiment of ResNet50 + DC + CA on INbreast Dataset (Accuracy in %)

BI-RADS	I	II	III	IV
I	95.8	4.2	0	0
II	11.1	69.4	16.7	2.8
III	11.1	27.8	61.1	0
IV	0	16.6	16.7	66.7

the MLO-view mammograms are redundant in this task or a more elegant network design is needed to utilize the MLO-view mammograms in an effective manner, which we will investigate further in the following work.

Different imaging systems or experimental settings generate images of different standards. A trained CNN can only properly handle the domain-specific images. Though including different types of images into the training process can help build a more robust CNN model, it is not realistic to collect a dataset which considers all the different possibilities. Thanks to the large capacity of CNNs, our classification model could be easily extended depending on the application situations. If the dataset to be processed is in a similar domain as the original dataset, the trained CNN model could be used directly. However, if the new dataset is in a very different domain from the original dataset, fine-tuning of the trained CNN is required before it could be successfully applied. Compared with training from scratch, fine-tuning of CNNs requires much fewer samples and the training process is significantly faster.

Our residual learning-based CNN model could serve as a baseline for mammographic density classification. In the future, we expect to collect more data to train a more powerful CNN model. We will go deeper into the class imbalance issue. As discussed in the results section, our architecture still needs to be optimized. We also plan to test the fine-tuning performance of the baseline model by using datasets that come from different systems or different experimental settings. We make our code and trained models publicly available to foster the research in the field.

6 CONCLUSION

In this study, we investigated the utilization of a radiomics-based method through residual learning for mammographic breast density classification. To elevate the classification performance of the networks, we integrated dilated convolutions and attention mechanisms into the naive deep residual networks. The proposed method is trained end-to-end and sophisticated pre-processing of the mammographic images, such as segmentation of the breast tissues, is not required. Our proposed method achieved superior classification performance on the clinical dataset, where a larger quantity of images were available and the image distributions over the different categories were relatively balanced. With this dataset, we also validated that multi-view inputs were beneficial to the breast density classification task. For the publicly available dataset, the proposed network still showed better performance than the naive residual networks but with decreased classification accuracy than the clinical dataset. Following this work, we plan to solve the data imbalance problem as well as the issues on how to properly train deep CNNs with small datasets and network architecture optimization.

REFERENCES

[1] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2017," *CA. Cancer J. Clinicians*, vol. 67, no. 1, pp. 7–30, 2017.

[2] F. Bray, P. Mccarron, and D. M. Parkin, "The changing global patterns of female breast cancer incidence and mortality," *Breast Cancer Res.*, vol. 6, no. 6, pp. 229–239, 2004.

[3] E. A. Sickles, "Breast cancer screening outcomes in women ages 40–49: Clinical experience with service screening using modern mammography," *J. Nat. Cancer Inst. Monographs*, vol. 22, no. 22, pp. 99–104, 1997.

[4] A. Oliver *et al.*, "A novel breast tissue density classification methodology," *IEEE Trans. Inf. Technol. Biomedicine*, vol. 12, no. 1, pp. 55–65, Jan. 2008.

[5] A. Oliver *et al.*, "Breast density analysis using an automatic density segmentation algorithm," *J. Digit. Imag.*, vol. 28, no. 5, pp. 604–612, 2015.

[6] A. Rampun, B. Scotney, P. Morrow, H. Wang, and J. Winder, "Breast density classification using local quinary patterns with various neighbourhood topologies," *J. Imag.*, vol. 4, no. 14, pp. 1–23, 2018.

[7] J. N. Wolfe, "Risk for breast cancer development determined by mammographic parenchymal pattern," *Cancer*, vol. 37, no. 5, pp. 2486–2492, 1976.

[8] N. F. Boyd *et al.*, "Quantitative classification of mammographic densities and breast cancer risk: Results from the canadian national breast screening study," *J. Nat. Cancer Inst.*, vol. 87, no. 9, pp. 670–675, 1995.

[9] M. Kallenberg *et al.*, "Unsupervised deep learning applied to breast density segmentation and mammographic risk scoring," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1322–1331, May 2016.

[10] C. W. Huo *et al.*, "Mammographic density—A review on the current understanding of its association with breast cancer," *Breast Cancer Res. Treatment*, vol. 144, no. 3, pp. 479–502, 2014.

[11] W. A. Berg, C. Campassi, P. Langenberg, and M. J. Sexton, "Breast imaging reporting and data system: Inter- and intraobserver variability in feature analysis and final assessment," *Amer. J. Roentgenol.*, vol. 174, no. 6, pp. 1769–1777, 2000.

[12] N. Karssemeijer, "Automated classification of parenchymal patterns in mammograms," *Phys. Medicine Biol.*, vol. 43, pp. 365–378, 1998.

[13] K. E. Martin *et al.*, "Mammographic density measured with quantitative computer-aided method: Comparison with radiologists' estimates and BI-RADS categories," *Radiology*, vol. 240, no. 3, pp. 656–665, 2006.

[14] C. Zhou *et al.*, "Computerized image analysis: Estimation of breast density on mammograms," *Med. Phys.*, vol. 28, no. 6, pp. 1056–1069, 2001.

[15] K. Bovis and S. Singh, "Classification of mammographic breast density using a combined classifier paradigm," in *Proc. 4th Int. Workshop Digit. Mammography*, 2002, pp. 177–180.

[16] N. M. Parthaláin, R. Jensen, Q. Shen, and R. Zwiggelaar, "Fuzzy-rough approaches for mammographic risk analysis," *Intell. Data Anal.*, vol. 14, pp. 225–244, 2010.

[17] Z. Chen, E. Denton, and R. Zwiggelaar, "Local feature based mammographic tissue pattern modelling and breast density classification," in *Proc. 4th Int. Conf. Biomed. Eng. Informat.*, 2011, pp. 351–355.

[18] I. Kumar, H. S. Bhadauria, and J. Virmani, "Wavelet packet texture descriptors based four-class BIRADS breast tissue density classification," *Procedia Comput. Sci.*, vol. 70, pp. 76–84, 2015.

[19] G. Litjens *et al.*, "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, 2017.

[20] Y. Zhou *et al.*, "A radiomics approach with CNN for shear-wave elastography breast tumor classification," *IEEE Trans. Biomed. Eng.*, vol. 65, no. 9, pp. 1935–1942, Sep. 2018.

[21] D. Shen, G. Wu, and H. I. Suk, "Deep learning in medical image analysis," *Annu. Rev. Biomed. Eng.*, vol. 19, no. 1, pp. 221–248, 2017.

[22] H. C. Shin *et al.*, "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1285–1298, May 2016.

[23] L. Zhang, L. Lu, R. M. Summers, E. Kebebew, and J. Yao, "Convolutional invasion and expansion networks for tumor growth prediction," *IEEE Trans. Med. Imag.*, vol. 37, no. 2, pp. 638–648, Feb. 2018.

[24] P. Lambin *et al.*, "Radiomics: Extracting more information from medical images using advanced feature analysis," *Eur. J. Cancer*, vol. 48, no. 4, pp. 441–446, 2012.

[25] R. J. Gillies, P. E. Kinahan, and H. Hricak, "Radiomics: Images are more than pictures, they are data," *Radiology*, vol. 278, no. 2, pp. 563–577, 2016.

[26] C. Jin, H. Cai, J. Wang, L. Li, W. Tan, and Y. Xi, "Discrimination of breast cancer with microcalcifications on mammography by deep learning," *Sci. Rep.*, vol. 6, 2016, Art. no. 27327.

[27] S. Suzuki *et al.*, "Mass detection using deep convolutional neural network for mammographic computer-aided diagnosis," in *Proc. 55th Annu. Conf. Soc. Instrum. Control Engineers Japan*, 2016, pp. 1382–1386.

[28] D. Wang, A. Khosla, R. Gargeya, H. Irshad, and A. H. Beck, "Deep learning for identifying metastatic breast cancer," 2016, *arXiv:1606.05718*. [Online]. Available: https://arxiv.org/abs/1606.05718

[29] L. Zhang, S. Jiang, Y. Zhao, J. Feng, B. W. Pogue, and K. D. Paulsen, "Direct regularization from co-registered contrast MRI improves image quality of MRI-guided near-infrared spectral tomography of breast lesions," *IEEE Trans. Med. Imag.*, vol. 37, no. 5, pp. 1247–1252, May 2018.

[30] G. Carneiro, J. Nascimento, and A. P. Bradley, "Automated analysis of unregistered multi-view mammograms with deep learning," *IEEE Trans. Med. Imag.*, vol. 36, no. 11, pp. 2355–2365, Nov. 2017.

[31] H. Sun *et al.*, "AUNet: Attention-guided dense upsampling networks for breast mass segmentation in whole mammograms," *Phys. Medicine Biol.*, to be published. [Online]. Available: https://iopscience.iop.org/article/10.1088/1361-6560/ab5745

[32] P. Fonseca *et al.*, "Automatic breast density classification using a convolutional neural network architecture search procedure," in *Proc. SPIE-The Int. Soc. Opt. Photon.*, 2015, Art. no. 941428.

[33] C. K. Ahn, C. Heo, H. Jin, and J. H. Kim, "A novel deep learning-based approach to high-accuracy breast density estimation in digital mammography," in *Proc. SPIE-The Int. Soc. Opt. Photon.*, 2017, Art. no. 101342O.

[34] S. Li, J. Wei *et al.*, "Computer-aided assessment of breast density: Comparison of supervised deep learning and feature-based statistical learning," *Phys. Medicine Biol.*, vol. 63, no. 2, 2018, Art. no. 025005.

[35] N. Wu *et al.*, "Breast density classification with deep convolutional neural networks," in *Proc. ICASSP-IEEE Int. Conf. Acoust. Speech Signal Process.*, 2018, pp. 6682–6686.

[36] 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," *Med. Phys.*, vol. 45, no. 1, pp. 314–321, 2018.

[37] M. Bahl *et al.*, "Mammographic breast density assessment using deep learning: Clinical implementation," *Radiology*, vol. 290, no. 1, pp. 52–58, 2018.

[38] T. Kyono, F. J. Gilbert, and M. V. D. Schaar, "MAMMO: A deep learning solution for facilitating radiologist-machine collaboration in breast cancer diagnosis," 2018, *arXiv:1811.02661*. [Online]. Available: https://arxiv.org/abs/1811.02661

[39] Z. Gandomkar, M. E. Suleiman, D. Demchig, P. C. Brennan, and M. F. McEntee, "BI-RADS density categorization using deep neural networks," in *Proc. SPIE-The Int. Soc. Opt. Photon.*, 2019, Art. no. 109520N.

[40] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *Proc. Int. Conf. Learn. Representations*, 2016, pp. 1–13. [Online]. Available: https://arxiv.org/abs/1511.07122

[41] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[42] F. Yu, V. Koltun, and T. Funkhouser, "Dilated residual networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017. pp. 472–480.

[43] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2921–2929.

[44] A. G. Roy, N. Navab, and C. Wachinger, "Concurrent spatial and channel 'Squeeze & Excitation' in fully convolutional networks," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, 2018, pp. 421–429.

[45] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.

[46] I. C. Moreira, I. Amaral, I. Domingues, A. Cardoso, M. J. Cardoso, and J. S. Cardoso, "INbreast: Toward a full-field digital mammographic database," *Acad. Radiol.*, vol. 19, no. 2, pp. 236–248, 2012.

[47] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[48] T. Elsken, J. H. Metzen, and F. Hutter, "Neural architecture search: A survey," 2018, *arXiv:1808.05377*. [Online]. Available: https://arxiv.org/abs/1808.05377

[49] R. Sun, X. Zhu, C. Wu, C. Huang, J. Shi, and L. Ma, "Not all areas are equal: Transfer learning for semantic segmentation via hierarchical region selection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4360–4369.

Cheng Li received her PhD degree from Nanyang Technological University, Singapore. She is a postdoctoral fellow in Paul C. Lauterbur Research Center for Biomedical Imaging, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China. Her research interests include medical image processing, signal processing, multi-modal image fusion, and machine learning.
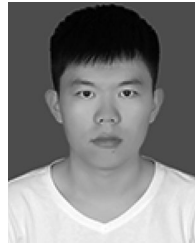
Jingxu Xu is currently working toward the master's degree in the School of Biomedical Engineering, Health Science Center, Shenzhen University, Shenzhen, China. His research interests include biomedical image analysis, deep learning, and software engineering.
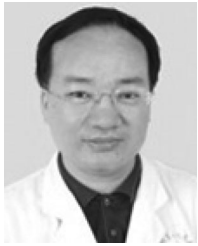
Qiegen Liu (Member, IEEE) received the PhD degree from Shanghai Jiao Tong University, Shanghai, China. He is an associate professor in Nanchang University, Nanchang, China. His research interests include dictionary learning, compressed sensing, image processing, and deep learning.

Yongjin Zhou received the PhD degree from Xi'an Jiaotong University, Xi'an, China. He is an associate professor in Shenzhen University, Shenzhen, China. His research interests include rehabilitation assessment with ultrasound and wearable health management.

Lisha Mou received the PhD degree from Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, China. She is an associate professor with the Second People's Hospital of Shenzhen. Her research interests include islet transplantation, the application of machine learning to medical imaging, and computer-aided diagnosis.

Zuhui Pu received the MD degree from the School of Medicine, Fudan University, Shanghai, China. He is a chief physician and professor with the Second People's Hospital of Shenzhen. His research contributions are in early diagnosis and treatment of coronary heart disease, screening methods for early lung cancer, and screening approaches for breast cancer.

Yong Xia received the PhD degree from Northwestern Polytechnical University, Xi'an, China. He is a professor with the Department of Computer Science, Northwestern Polytechnical University, Xi'an, China. His research contributions are in medical image analysis, image and multimedia information processing, computer-aided diagnosis, pattern recognition, and machine learning fields.

Hairong Zheng (Senior Member, IEEE) received the PhD degree from the University of Colorado, America. He is a professor in Paul C. Lauterbur Research Center for Biomedical Imaging, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China. He is the associate dean of Shenzhen Institutes of Advanced Technology. His research contributions are in developing techniques and instruments for medical imaging, including ultrasound elastography, magnetic resonance imaging, and molecular imaging.

Shanshan Wang (Senior Member, IEEE) received the dual PhD degree in information technologies and biomedical engineering from the University of Sydney, Sydney, Australia and Shanghai Jiao Tong University, Shanghai, China. She is an associate professor in Paul C. Lauterbur Research Center for Biomedical Imaging, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China. Her research interests include machine learning, fast medical imaging, and radiomics. She has published more than 40 journal and conference papers in these areas.

▷ **For more information on this or any other computing topic, please visit our Digital Library at** www.computer.org/csdl.