# High-Resolution Breast Cancer Screening with Multi-View Deep Convolutional Neural Networks

Krzysztof J. Geras[1,3], Stacey Wolfson[3], Yiqiu Shen[1], Nan Wu[1], S. Gene Kim[3,4], Eric Kim[3], Laura Heacock[3], Ujas Parikh[3], Linda Moy[3,4], Kyunghyun Cho[1,2,5]



***Abstract*—Advances in deep learning for natural images have prompted a surge of interest in applying similar techniques to medical images. The majority of the initial attempts focused on replacing the input of a deep convolutional neural network with a medical image, which does not take into consideration the fundamental differences between these two types of images. Specifically, fine details are necessary for detection in medical images, unlike in natural images where coarse structures matter most. This difference makes it inadequate to use the existing network architectures developed for natural images, because they work on heavily downscaled images to reduce the memory requirements. This hides details necessary to make accurate predictions. Additionally, a single exam in medical imaging often comes with a set of views which must be fused in order to reach a correct conclusion. In our work, we propose to use a multi-view deep convolutional neural network that handles a set of high-resolution medical images. We evaluate it on large-scale mammography-based breast cancer screening (BI-RADS prediction) using 886,000 images. We focus on investigating the impact of the training set size and image size on the prediction accuracy. Our results highlight that performance increases with the size of training set, and that the best performance can only be achieved using the original resolution. In the reader study, performed on a random subset of the test set, we confirmed the efficacy of our model, which achieved performance comparable to a committee of radiologists when presented with the same data.**

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

## I. Introduction

Breast cancer is the second leading cancer-related cause of death among women in the United States. It is estimated that 232,000 women were diagnosed with breast cancer and approximately 40,000 died from the disease in 2015 [1]. Screening mammography is the main imaging test used to detect occult breast cancer. Multiple randomized clinical trials have shown a 30% reduction in mortality in asymptomatic women who were undergoing screening mammography [2], [3]. Although mammography is the only imaging test that reduced breast cancer mortality [2], [3], [4], the appropriate screening interval for mammograms has been the subject of public debate with different professional societies offering varying guidelines for mammographic screening [2], [3], [4], [5]. In particular, there has been public discussion regarding the potential harms of screening. These harms include false positive recalls and false positive biopsies as well as anxiety caused by recall for diagnostic testing after a screening exam. Overall, the recall rate following a screening mammogram is between 10-15%. This equates to about 3.3 to 4.5 million callback exams for additional testing [6].

The vast majority of the women asked to return following an inconclusive mammogram undergo another mammogram and/or ultrasound for clarification. Most of these false positive findings are found to represent normal breast tissue. Only 10% to 20% of women who have an abnormal screening mammogram are recommended to undergo a biopsy. Only 20-40% of these biopsies yield a diagnosis of cancer [7]. In 2014, over 39 million screening and diagnostic mammography exams were performed in the US. Therefore, in addition to the anxiety from undergoing a false positive mammogram, there are significant costs associated with unnecessary follow ups and biopsies. Clearly, there is an unmet need to shift the balance of routine breast cancer screening towards more benefit and less harm.

### A. Breast Cancer Screening as a Deep Learning Task

Deep learning has recently seen enormous success in challenging problems such as object recognition in natural images, automatic speech recognition and machine translation [8]. This success has prompted a surge of interest in applying deep convolutional networks (DCN) to medical imaging. Many recent studies have shown the potential of applying such networks to medical imaging, including breast screening mammography; however, without investigating the fundamental differences between medical and natural images and their impact on the design choices and performance of proposed models. For instance, many recent works have either significantly downscaled a whole image or focused on classifying a small region of interest. This might be detrimental to performance of such models given the well-known dependency of breast cancer screening on fine details, such as the existence of a cluster of microcalcifications, as well as global structures, for example the symmetry between two breasts. Furthermore, the potential of DCNs has only been assessed in limited settings of small data sets often consisting of less than one thousand images, while the success of such networks in natural

[1]Center for Data Science, New York University, 60 5th Ave, New York, NY 10011

[2]Courant Institute of Mathematical Sciences, New York University, 251 Mercer St, New York, NY 10012

[3]Center for Biomedical Imaging, Radiology, NYU School of Medicine, 660 1st Avenue, New York, NY 10016

[4]Perlmutter Cancer Center, NYU Langone Medical Center, 160 E 34th St, New York, NY 10016

[5]CIFAR Azrieli Global Scholar

object recognition is largely attributed to the availability of more than one million annotated images. This further hinders our understanding of the true potential of DCNs in medical imaging, particularly in breast cancer screening.

In this work, we conducted an investigation into analyzing and understanding fundamental properties of deep convolutional networks in the context of breast cancer screening. We started by building a large-scale data set of 201,698 screening mammographic exams (886,437 images) collected at multiple sites of our institution. We developed a novel DCN that is able to handle multiple views of screening mammography and to utilize large high-resolution images without downscaling. We refer to this DCN as a multi-view deep convolutional network (MV-DCN). Our network learns to predict the assessment of a radiologist, classifying an incoming example as BI-RADS 0 ("incomplete"), BI-RADS 1 ("normal") or BI-RADS 2 ("benign finding"). We studied the impact of the data set size and image resolution on the screening performance of the proposed MV-DCN, which would serve as a *de facto* guideline for optimizing future deep neural networks for medical imaging. We further investigated the potential of the proposed MV-DCN by visualizing its predictions. Finally, we conducted a reader study, which showed that our model, on a random subset of the test set, is almost as accurate as a committee of radiologists presented with the same data. Furthermore, we found that we obtain the best results by averaging the predictions of our model's with the predictions of the committee of the radiologist.

## II. High-Resolution Multi-View Deep Convolutional Neural Networks

### A. Deep Convolutional Neural Network

A deep convolutional neural network [9], [10] is a classifier that takes an image $\mathbf{x}$ as input, often with multiple channels corresponding to different colors (e.g., RGB), and outputs the conditional probability distribution over the categories $p(y|\mathbf{x})$. This is done by a series of nonlinear functions that gradually transform the input pixel-level image. A major property of the deep convolutional network, which distinguishes it from a multi-layer perceptron, is that it heavily relies on convolutional and pooling layers, which make the network invariant to local translation of visual features in the input.

### B. Multi-View Deep Convolutional Neural Network

Object recognition tasks with natural images usually involve only one object at a time, in contrast an exam in medical imaging often comes with a set of views. For instance, it is standard in screening mammography to obtain cranial caudal (CC) and mediolateral oblique (MLO) views for each breast of a patient, resulting in a set of four images. We will refer to them as L-CC, R-CC, L-MLO and R-MLO (Figure 1).

There is a rich literature on building deep neural networks for multi-view data. Most of them fall into one of two families. First, there are works on unsupervised feature extraction from multiple views using a variant of deep autoencoders [11], [12], [13]. They usually train a multi-view deep neural network with unlabeled examples, and use the output of such a network as a feature extractor, followed by a standard classifier. On the other hand, Su et al. [14] proposed to build a multi-view deep convolutional network directly for classification.

We propose a variant of MV-DCN which was motivated by Su et al. [14]. This MV-DCN computes the output in two stages. In the first stage, a number of convolutional and pooling layers is separately applied to each of the views. We denote such view-specific representation by $\mathbf{h}_{\mathrm{v}}$, where v refers to the index of the view. These view-specific representations are concatenated to form a vector, $[\mathbf{h}_{\mathrm{L-CC}}, \mathbf{h}_{\mathrm{R-CC}}, \mathbf{h}_{\mathrm{L-MLO}}, \mathbf{h}_{\mathrm{R-MLO}}]$, which is an input to the second stage - a fully connected layer followed by a softmax layer producing output distribution $p(\mathbf{y}|x)$.

The whole network is trained jointly by stochastic gradient descent with backpropagation [15]. Furthermore, we employ a number of regularization techniques to avoid the behavior of overfitting due to the relatively small size of training dataset, such as data augmentation by random cropping [16] and dropout [17]. These will be described later in detail.

### C. High-Resolution Convolutional Neural Network

It is common in object recognition and detection in natural images to heavily downscale an original high-resolution image. For instance, the input to the network of the best performer in ImageNet Challenge 2015 (classification task) was an image downscaled to $224 \times 224$ [18]. This is often done to improve the computational efficiency, both in terms of computation and memory, and also because no significant improvement has been observed with higher-resolution images. It reflects an inherent property of natural images, in which the objects of interest are usually presented in relatively larger portions than other objects and what matters most are their macro-structures, such as shapes, colors and other global descriptors. However, downscaling of an input image is not desirable in the case of medical images, and in particular for early-stage screening based on breast mammography. Often a cue for diagnosis is a subtle finding which may be identified only at the original resolution.

In order to address the computational issues of handling full-resolution images, we propose to use aggressive convolution and pooling layers. First, we use convolution layers with strides larger than one in the first two convolutional layers. Also, the first pooling layer has a larger stride than the other pooling layers. As a result of this, we greatly reduce the size of feature maps early in the network. Although this aggressive convolution and pooling loses some spatial information, the parameters of the network are adjusted to minimize this information loss during training. This is unlike downscaling of the input, which loses information unconditionally. Second, we average feature maps in the last layer before concatenating them [19], instead of simply flattening the feature maps and then concatenating them [16], [20]. This drastically reduces the dimensionality of the view-specific vector without much, if any, performance degradation [21]. Using both of these approaches, we are able to build an MV-DCN that takes four $2600 \times 2000$ pixels images (one per view) as input without any downscaling.

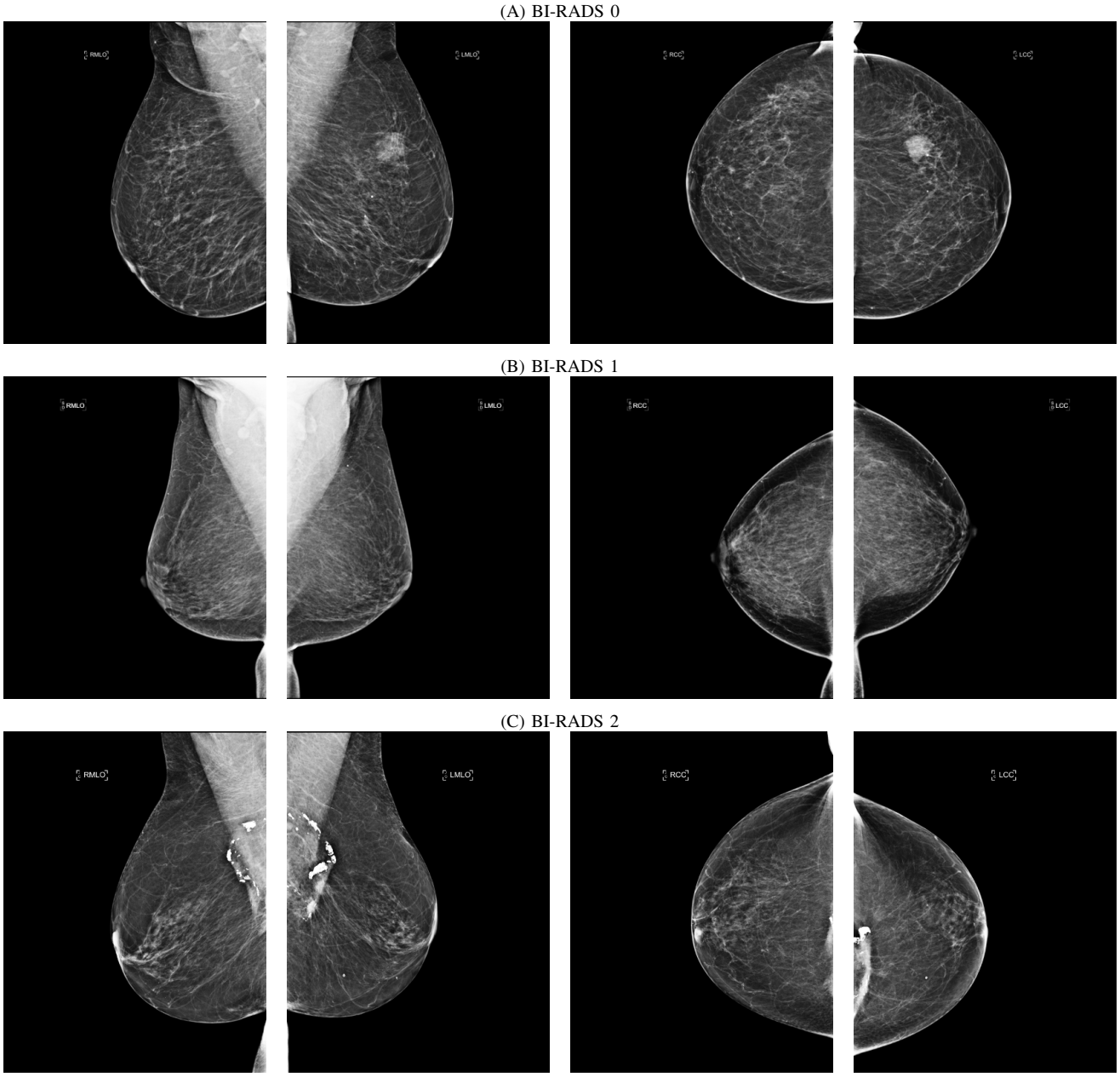



Fig. 1. The four standard views used in our experiments for exams categorized as BI-RADS 0 (A), BI-RADS 1 (B) and BI-RADS 2 (C). (A) In the left breast, there is a round mass with irregular margins. The patient was recalled, because additional mammographic views and a breast ultrasound were necessary to further characterize this mass. This mass turned out to be an invasive ductal carcinoma. (B) This is a normal mammogram in a patient with a scattered fibroglandular tissue breast pattern. No abnormalities were seen. (C) This patient has a scattered fibroglandular tissue breast pattern. In the posterior depth of the both breasts, near the chest wall, there are calcified masses consistent with post-surgical changes.

## III. Related Work

Let us briefly review recent deep learning based approaches to breast mammography, summarized in Table I.

*a) Multi-Stage vs. End-to-End Approaches.:* Traditionally breast cancer screening and lesion detection are done in three stages: detection, analysis and final assessment/management. In the first stage, a breast mammography image is segmented into different types of regions, such as foreground (breast) and background. Within the segmented region of breast, the second stage focuses on extracting a set of regions of interest (ROI) that will be examined in more detail. In the third stage, each of those ROI's is determined to be a malignant lesion or not. The outcome of the third stage is used to make the final decision on a given case consisting of multiple views.

Most of the recent research on applying deep learning to breast mammography has focused on replacing one or more stages in this existing multi-stage pipeline; for instance, mass detection [22], [23], [24]. In their work, a deep neural network is trained to determine whether a small patch is a

mass. Others have focused on training a deep neural network for classifying a small region of interest into one of a few categories, assuming an existing mass detection system [25], [26], [27], [28].

On the other hand, a small number of research groups have considered replacing the whole multi-stage approach with a single, or a series of, trainable machine learning algorithms. Kooi et al. [29] proposed to use a random forest classifier for mass detection followed by a DCN that classifies each detected mass. A similar approach was proposed by Becker et al. [30]. Akselrod-Ballin et al. [31] further proposed to use deep convolutional networks for both mass detection and classification, potentially enabling end-to-end training. Two groups [32], [33] went even further by proposing a single deep convolutional network that classifies a whole image, or a set of multiple views. The work by Carneiro et al. [33] is closest to our approach in this paper. In both works, a single deep convolutional neural network takes as input a set of multiple views of an exam and predicts its BI-RADS label.

*b) Data Size.:* Although it is recognized that one of the driving forces behind the success of deep learning is the availability of large scale data, it has not been exploited when applying deep learning to mammography. As evident in Table I, most of the recent works use less than 1,000 images for both training and testing. To avoid the issue of small training data, most of the earlier works resorted to training with many small patches, or ROI's, avoiding end-to-end training. One exception is the work of Carneiro et al. [33] in which they use the whole image with, however, the deep convolutional network pretrained for object recognition in natural images. Unlike these earlier approaches, we use a large-scale data set of an unprecedented size, consisting of 886,437 images. This allows us to carefully study the impact of the size of training data set.

*c) Natural vs. Controlled Distribution.:* Breast screening is aimed at a general population rather than a selected group of patients. This implies that the distribution of the screening outcome is heavily skewed toward "normal" (BI-RADS 1). In our training set which closely follows a general population distribution, approximately 46% of the cases were assigned BI-RADS 1 ("normal"), while 41% were assigned BI-RADS 2 ("benign finding") and 13% BI-RADS 0 ("incomplete"). This is in contrast to two widely-used, publicly available datasets, INBreast [34], DDSM [35], [36] and other curated small-scale datasets from recent literature (see those in Table I). These datasets are often constructed to include approximately the same proportions of normal and abnormal cases, resulting in, what we refer to as, a *controlled distribution* of outcomes which differs from a *natural distribution*. For instance, INBreast has approximately achieved a balance between benign and malignant cases. This type of artificial balancing, or equivalently upsampling of malignant cases, may bias a model to more often predict a given case as malignant and require a recall more often than necessary. Unlike these earlier works, in this paper, we use the full data without artificial balancing of outcomes to ensure that any trained deep convolutional network will closely reflect the natural distribution of outcomes.

TABLE I
PREVIOUS WORKS ON DEEP LEARNING FOR BREAST MAMMOGRAPHY.

| task□ | ref. | E2E• | #images† | image size‡ | MV♡ | input♠ | dist.○ |
|---|---|---|---|---|---|---|---|
| BI-RADS | ⋆ | √ | 829k (57k) | 2600×2000 | √ | IMG | N |
| | [33] | √ | 680 (≈ 340) | 264×264 | √ | IMG | C |
| | [32] | √ | 410 (CV) | 224×224 | √ | IMG | C |
| | [31] | √ | 850 (≈ 170) | 800×800 | | IMG | C |
| lesion | [25] | | 607 (CV) | 512×512 | | ROI | C |
| | [29] | | 44,000 (18,000) | 250×250 | | ROI | N |
| | [26] | | 1820 (182) | 224×224 | | ROI | C |
| | [27] | | 736 (≈300) | 150×150 | | ROI | C |
| | [28] | | 1606 (≈378) | 13×13 | √ | ROI | N |
| mass | [22] | | 116 (CV) | 32×32 | | ROI | C |
| | [23] | | 410 (CV) | 264×264 | | ROI | C |
| | [24] | √ | 2500 (250) | 256×256 | | ROI | C |
| MC | [37] | | 1000 (204) | N/A♣ | | ROI | C |
| | [38] | | 1410 (N/A) | N/A♣ | | ROI | C |

When more than one data set was used, we list the size of the largest data set. ⋆ denotes this paper. The table should be read with the following footnotes. □ The target task; BI-RADS: BI-RADS prediction, lesion: lesion classification (benign vs. malignant), mass: mass detection, and MC: microcalcification detection. • Whether the proposed system is trainable end-to-end. For instance, a system that requires an external system for extracting regions of interest (ROI) is not end-to-end, while a system that uses convolutional networks for both ROI extraction and lesion classification is. † In the parentheses there is the number of test images or "CV" if cross-validation is used. ‡ denotes the size of the input image to a deep neural network. ♡ Whether multiple views per one exam are utilized. ○ Whether the data reflects natural distribution (N) or controlled distribution (C). ♠ Whether the input to a deep neural network is a whole image (IMG) or a small subset (ROI). ♣ Did not use images as input to the learning algorithm.

## IV. DATA

### A. Collection

This is a Health Insurance Portability and Accountability (HIPAA)-compliant, retrospective study approved by our Institutional Review Board. Consecutive screening mammograms for 129,208 patients aged[1] between 19 and 99 (mean: 57.2, std: 11.6) collected within seven years (2010-2016) at five imaging sites affiliated with New York University School of Medicine were used in this study. These imaging centers are located in the New York City metropolitan area (a large academic center and two large ambulatory care practices), where, altogether, over 70,000 mammograms are performed annually. The ethnic makeup of the patient cohort for this study reflects the population pool in NYC, which is 50% Caucasian, 30% African American, 5% Asian and 15% Hispanic.

### B. Data Statistics

We used all data that we were able to collect and did not exclude any data unless they were acquired incorrectly[2]. We divided the data into disjoint training, validation and test sets in the following manner. To start with we sorted all patients

[1]When more than one exam for a patient was in the data set, we included the ages of that patient at the time of all exams to compute the values above.

[2]We only excluded images if they were of views which should not be taken in screening mammography, if they were technically not correct (e.g. if they did not have a time stamp), if they were smaller than 2600 × 2000 pixels in either of the corresponding dimensions or if their magnification factor was smaller than 1.0 or bigger than 1.1. We used all exams that had at least one correct image for each of the standard views.

according to the date of their latest exam in the data set. We use the first 80% of the patients in this order as the training data, the next 10% as the validation data and the last 10% as the test data. For each patient in the test set, we evaluate our model's performance in predicting only the label for the latest exam of each patient. This way we can reliably estimate the level of accuracy we would achieve if we tested our model on future exams. There are altogether 129,208 patients, 201,698 exams and 886,437 images in the data set.

TABLE II
DISTRIBUTION OF DATA ASSOCIATED WITH DIFFERENT BI-RADS IN TRAINING, VALIDATION AND TEST DATA. EACH CELL IN THE TABLE HAS THE FOLLOWING FORMAT: NUMBER OF EXAMS / NUMBER OF IMAGES.

| | BI-RADS 0 | BI-RADS 1 | BI-RADS 2 |
|---|---|---|---|
| Training | 21946 / 95471 | 74832 / 327035 | 67446 / 298680 |
| Validation | 2634 / 11471 | 11542 / 50627 | 10376 / 46178 |
| Test | 1341 / 5871 | 5986 / 26213 | 5595 / 24891 |

### C. Data preprocessing and augmentation

We normalized the images in the following way. For each image we computed the mean, $\mu$, and the standard deviation, $\sigma$, of its pixels. We then subtracted $\mu$ from each pixel and divided each pixel by $\sigma$. Additionally, we flipped horizontally the images of R-CC and R-MLO views so that the breast was always on the same side of the image.

Since the images vary in size and a large fraction of the surface of each image is empty, we cropped all of them to the size of $2600 \times 2000$ pixels. We did it for two reasons. First, to unify the sizes of the images (which we need to put them in mini-batches during training) while keeping them at a similar scale and, second, to avoid processing the background which does not contain any information. The position of the crop was determined in the following manner. First, the crop area was placed leftmost on the horizontal axis and centrally on the vertical axis. To augment the data set, noise was added to this position. Let us denote the number of pixels between the top border of the crop area and the top border of the image by $b_{\text{top}}$ and analogously define $b_{\text{bottom}}$ and $b_{\text{right}}$. We drew a number, $t_{\text{vertical}}$ from a uniform distribution $\mathcal{U}(-\min(b_{\text{top}}, 100), \min(b_{\text{bottom}}, 100))$ and $t_{\text{horizontal}}$ from $\mathcal{U}(0, \min(b_{\text{right}}, 100))$. Finally we translated the position of the crop area by $t_{\text{horizontal}}$ pixels horizontally and $t_{\text{vertical}}$ pixels vertically. During training this noise was sampled independently every time an image is used. During validation there was no noise added to the position of the crop area. At test time, we fed ten sets of four randomly cropped views to the network. The final prediction was made by averaging predictions for all crops. The aim of this averaging is twofold; first, to use information from outside the center of the image while keeping the size of the input fixed and second, to make prediction of the network more stable. A small fraction of data contains more than one image per view. For such cases one image per view was sampled randomly and uniformly each time an exam was used during training and testing. During validation the image with the earliest time stamp was always used.

| layer | kernel size | stride | #maps | repetition |
|---|---|---|---|---|
| global average pooling | | | 256 | |
| convolution | 3×3 | 1×1 | 256 | ×3 |
| max pooling | 2×2 | 2×2 | 128 | |
| convolution | 3×3 | 1×1 | 128 | × 3 |
| max pooling | 2×2 | 2×2 | 128 | |
| convolution | 3×3 | 1×1 | 128 | × 3 |
| max pooling | 2×2 | 2×2 | 64 | |
| convolution | 3×3 | 1×1 | 64 | × 2 |
| convolution | 3×3 | 2×2 | 64 | |
| max pooling | 3×3 | 3×3 | 32 | |
| convolution | 3×3 | 2×2 | 32 | |
| input | | | 1 | |

Fig. 2. Description of one deep convolutional network column for a single view. It transforms the input view (a gray-scale image) into a 256-dimensional vector.

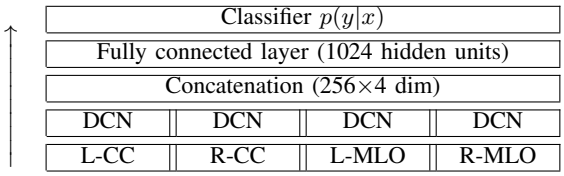



Fig. 3. An overview of the proposed multi-view deep convolutional network. DCN refers to the convolutional network network column from Figure 2. The arrow indicates the direction of information flow.

## V. SETTINGS

### A. Evaluation Metrics

When there are two classes the most frequently applied performance metric is the AUC (area under the ROC curve). However, since there are three classes in our learning task, we cannot apply this metric directly. Instead we computed three AUCs, each time treating one of the three classes as a positive class and the remaining two as negative. We used the macro average of the three AUCs, abbreviated as macAUC, as the main performance metric in this work.

Unlike other widely used nonlinear classifiers, such as a support vector machine or a random forest, a deep convolutional neural network outputs a proper conditional distribution $p(y|\mathbf{x})$. It allows us to compute the network's confidence in its prediction by computing the entropy of this distribution, i.e.,

$$H(y|\mathbf{x}) = -\sum_{y' \in \mathcal{C}} p(y'|\mathbf{x}) \log p(y'|\mathbf{x}), \tag{1}$$

where $y'$ iterates over all possible classes $\mathcal{C}$. The larger the entropy, the less confident the network is about its prediction. Based on $H$, we can quantify the change in accuracy (measured by AUC) with respect to the network's confidence.

### B. Model Setup

The overall architecture of our network is shown in Figure 3. Each column corresponding to a different view has an architecture described in Figure 2. We applied the rectifier function after convolutional layers. In addition to augmenting the data by cropping the images at random positions, we regularized the network in three ways. First, we tied the weights in the corresponding columns, i.e., the parameters of the columns

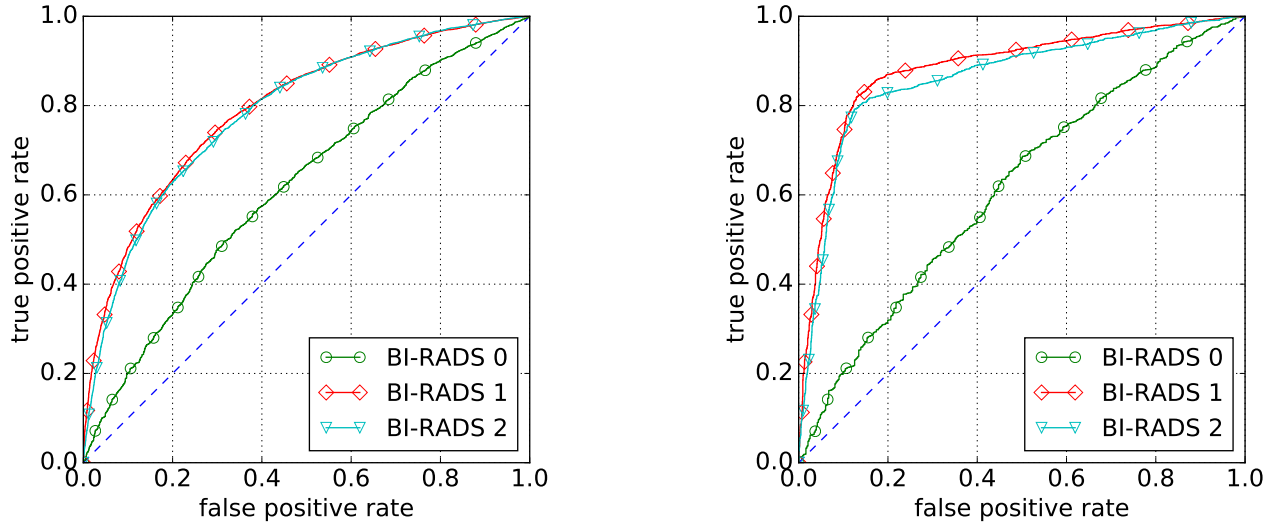



Fig. 4. ROCs computed with all test data (left) and ROCs computed with test data which the network was confident about (right). ROCs for BI-RADS 1 and BI-RADS 2 classes improve a lot for confident examples while BI-RADS 0 remain similar.

processing L-CC and R-CC views were shared as were those of the columns processing L-MLO and R-MLO views. Second, we added Gaussian noise to the input (with the mean of zero and the standard deviation of 0.01). Third, we applied dropout (with a rate of 0.2) after the fully connected layer. We turned off the input noise and dropout during validation and testing.

The parameters of the network were initialized using the recipe of Glorot & Bengio [39] and learned using the Adam algorithm [40] with the initial learning rate of $10^{-5}$. Due to the memory limitations of our hardware, the mini-batch size was set to four. We trained the network for up to 100 epochs, which takes approximately four weeks using one NVIDIA Tesla V100 GPU. After each training epoch we computed the macAUC on the validation set. We reported the test error of the model which achieved the lowest macAUC on the validation set.

We made the code allowing to run our best network available online at https://github.com/nyukat/BIRADS_classifier.

## VI. Quantitative Results Analysis

### A. Effect of Scale

First, we validated our earlier claim on the need of large-scale data for harnessing the most out of deep convolutional neural networks. We trained separate networks on the training sets of different sizes; 100%, 50%, 20% and 10%, 5%, 2% and 1% of the original training set[3]. In Table III, we observed that the classification performance improves as the number of training examples increases. This shows the importance of using a large training set. This is consistent with observations made in many other fields such as computer vision, natural language processing and speech recognition [8].

[3]We created the subsets of the original training set by random sampling without replacement.

### B. Effect of Resolution

We then investigated the effect of resolution of input images. Using the full training set, we trained networks with varying input resolutions; scaling both dimensions of the input by $\times 1/8$, $\times 1/4$ and $\times 1/2$. We used bicubic interpolation to downscale the input. When the input resolution is significantly smaller than the original some convolutional layers in the later stages cannot be applied because the size of the feature maps becomes smaller than the size of a convolutional kernel. In that case, we simply skipped the remaining layers until the global average pooling. As shown in Table IV, we already saw a drop in performance when each dimension of the input was downscaled by half. Further degradation of performance was observed with more aggressive downscaling.

### C. Confidence

Additionally, we checked how our model is performing for test examples depending on how confident it is about its predictions. We measure confidence of predictions in terms of the entropy of the output distribution (cf. Equation 1). This is how we performed the procedure allowing us to quantify this property of our model. First, we divided the exams in the validation set between the three classes. For each class separately we sorted the exams according to the entropy of predictions made by our model. Lets define (for each class

TABLE III
The effect of changing the fraction of the training data used. Increasing the amount of data yields better results.

| fraction | 1% | 2% | 5% | 10% | 20% | 50% | 100% |
|---|---|---|---|---|---|---|---|
| 0 vs. others | 0.541 | 0.550 | 0.559 | 0.564 | 0.570 | 0.604 | 0.618 |
| 1 vs. others | 0.534 | 0.631 | 0.707 | 0.738 | 0.749 | 0.774 | 0.794 |
| 2 vs. others | 0.537 | 0.628 | 0.715 | 0.742 | 0.752 | 0.771 | 0.787 |
| macAUC | 0.537 | 0.603 | 0.660 | 0.681 | 0.690 | 0.716 | 0.733 |
| HC-macAUC | 0.554 | 0.652 | 0.710 | 0.751 | 0.744 | 0.778 | 0.787 |

TABLE IV
THE EFFECT OF DECREASING THE RESOLUTION OF THE IMAGE.

| scale | ×1/8 | ×1/4 | ×1/2 | ×1 |
|---|---|---|---|---|
| 0 vs. others | 0.587 | 0.585 | 0.611 | 0.618 |
| 1 vs. others | 0.718 | 0.742 | 0.779 | 0.794 |
| 2 vs. others | 0.729 | 0.750 | 0.777 | 0.787 |
| macAUC | 0.678 | 0.692 | 0.722 | 0.733 |
| HC-macAUC | 0.743 | 0.753 | 0.782 | 0.787 |

separately) $t_k$ as the threshold such that $k$ percent of the examples in the validation set have entropy (of the predictions made by our model) smaller than $t_k$. Then, for the examples from the test set (again, for each class separately) and we selected only those for which the entropy (of the prediction of our model) was lower than $t_k$. For these examples, we re-computed AUCs and macAUC. When k = 30 we call macAUC computed for this subset of data the high confidence macAUC (HC-macAUC). As shown in Table V, confident predictions of the proposed model are more accurate. This phenomenon was apparent in in all the experiments (see Table III, Table IV and Figure 4).

TABLE V
AVERAGE AUC (MACAUC) AS A FUNCTION OF THE CONFIDENCE THRESHOLD $T_{P\%}$. WHEN $P = 30\%$, WE REFER TO THE MACAUC AS A HIGH-CONFIDENCE MACAUC (HC-MACAUC).

| $T_{P\%}$ | $T_{10\%}$ | $T_{20\%}$ | $\boldsymbol{T_{30\%}}$ | $T_{50\%}$ | $T_{100\%}$ |
|---|---|---|---|---|---|
| macAUC | 0.865 | 0.827 | 0.811 | 0.781 | 0.732 |

## VII. VISUALIZATION

A flip side of high effectiveness of a deep convolutional neural network is the difficulty in interpreting its internal processing. Only recently there have been some efforts on visualizing deep convolutional neural networks for computer vision [41], [42]. These recent approaches, however, are not computationally efficient and are not easy to apply to medical images for a number of reasons, including the need for training with a large data set [41] and the availability of good image statistics [42]. Instead, we propose a simpler visualization technique in this paper that does not require any further training.

We look at the sensitivity of the network's output to the perturbation of each input pixel. The network outputs the conditional distribution over all the categories, and we can measure the entropy (or confidence) of the predictive distribution $\mathcal{H}(y|\mathbf{x})$. We can use standard backpropagation to compute $\left|\frac{\partial \mathcal{H}}{\partial \mathbf{x}^v_{ij}}\right|$ for the pixel $(i, j)$ of the $v$-th view. Those input pixels that influence the confidence of the network will have high values, and those that do not contribute much will have low values ($\approx 0$). We show two examples of such visualization for patients which were confirmed by a follow-up examination to have breast cancer in Figure 5.

## VIII. READER STUDY

To understand the limit of performance possible to achieve on this task, we conducted a reader study with four human experts, who all were doctors experienced in reading breast cancer screening exams. The experts were all shown the same 500 exams randomly drawn from the test set, each with at least four images corresponding to the standard views used in screening mammography. For each exam, they were asked to indicate the most likely BI-RADS label according to their judgement. We first measured agreement between the radiologists themselves and between the radiologists and the labels in the data. The results are shown in Table VI. We can clearly observe that the agreement between different radiologist as well as between radiologists and the labels in the data is low. To obtain probabilistic predictions from this group of experts, we represented their classifications as one-hot vectors and averaged them for each exam. On the random subset of data used in our reader study (which turned out to be a difficult subset, cf. Table III) such a committee of radiologists achieved the macUAC of 0.704, while our model achieved the macUAC of 0.688. We conclude from these results that predicting BI-RADS without prior exams and information about the patient is very difficult even for well-trained human experts. Our neural network is already performing well in comparison. It is interesting to note that our model is clearly worse than the committee of the radiologists in recognizing BI-RADS 0 and clearly better in recognizing BI-RADS 2. We also evaluated an ensemble, created by equally weighting predictions of the committee of radiologists and our neural network (Table VI). For each of the BI-RADS categories, the ensemble was at least as good as any of its two base elements.

TABLE VI
RESULTS OF OUR READER STUDY COMPARING ACCURACIES OBTAINED BY THE COMMITTEE OF RADIOLOGISTS, OUR NEURAL NETWORK (MV-DCN) AND AN ENSEMBLE OF THE TWO.

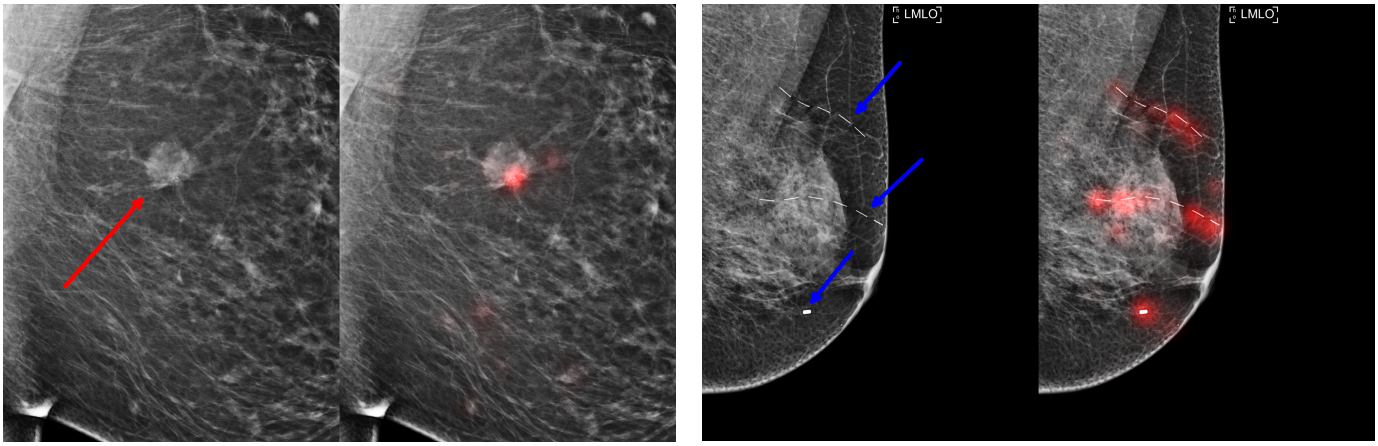
| | radiologists | MV-DCN | radiologists + MV-DCN |
|---|---|---|---|
| 0 vs. others | 0.650 | 0.547 | 0.653 |
| 1 vs. others | 0.765 | 0.757 | 0.792 |
| 2 vs. others | 0.699 | 0.759 | 0.759 |
| macAUC | 0.704 | 0.688 | 0.735 |

TABLE VII
AGREEMENT (COHEN'S KAPPA) IN BI-RADS CATEGORIZATION BETWEEN DIFFERENT RADIOLOGISTS (R1, R2, R3, R4) AND LABELS IN THE DATA SET (L).

| | L | R1 | R2 | R3 | R4 |
|---|---|---|---|---|---|
| L | | 0.29 | 0.24 | 0.24 | 0.26 |
| R1 | | | 0.29 | 0.34 | 0.35 |
| R2 | | | | 0.48 | 0.45 |
| R3 | | | | | 0.50 |
| R4 | | | | | |

## IX. CONCLUSIONS

In this paper we have made a first step towards end-to-end large scale training of multi-view deep convolutional networks for breast cancer screening. We have shown experimentally that it is essential to keep the images at high-resolution. We expect this to hold for other learning tasks with medical images where fine details determine the outcome. We also demonstrated it is necessary to use a large number of exams. Although we used the largest breast cancer screening data set




(a) BI-RADS 0 (b) BI-RADS 2

Fig. 5. Examples of visualizations of decisions made by our model. On the left in both (a) and (b) panels there are images of breast with arrows indicating possible suspicious findings. On the right in both (a) and (b) panels there are the same images as on the left in the corresponding panels with regions of the images (highlighted in red) which influence confidence of predictions of our neural network. Please note that our visualization highlights parts of the image that are relevant for all classes (BI-RADS 0, BI-RADS 1, and BI-RADS 2) and that those highlighted areas include locations indicated in the images on the left. Panel (a) shows the right breast of a 61 years old patient who was assigned BI-RADS 0 in her screening mammography. Biopsy confirmed that the finding indicated in the image by the right arrow was invasive ductal carcinoma. Panel (b) shows the left breast of a 62 years old patient. She had a prior breast surgery and a biopsy marker in one of her breasts as indicated by the blue arrows. Our neural network correctly and confidently predicted that the artifacts were benign and indicated scar markers and a biopsy clip.

ever reported in literature, the performance of our model has not saturated and is expected to improve with more data.

Our network's performance, just like performance of the doctors participating in our reader study, was lowest on differentiating BI-RADS 0 from the other classes. Doctors often disagree on how a particular exam should be classified [3] and in fact, less than 1% of the screening population has cancer [2], [5]. We expect that this problem can be alleviated by using instead the information on whether a person actually went on to develop breast cancer in the future as a label.

It is also worth noting that, because of limited computational resources, we had to heavily rely on our experience in the choice of learning hyperparameters. We did not perform a systematic search for optimal hyperparameters, which often has a great impact on the performance of a neural network in limited data scenarios [43], [44]. The methods we used in this work are powerful and our results can be improved simply by the means of applying more computational resources without significantly changing the methodology.

## Acknowledgments

We would like to thank Jure Žbontar, Yann LeCun, Pablo Sprechmann, Cem Deniz, Jingyi Su and Masha Zorin for insightful comments on this work, as well as Jason Phang and Jungkyu Park for creating a PyTorch version of the code. We would also like to thank Joe Katsnelson and Mario Videna for their efforts in supporting our computing environment.

## References

[1] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2015," *CA: a cancer journal for clinicians*, vol. 65, no. 1, pp. 5–29, 2015.

[2] S. W. Duffy, L. Tabar, H. H. Chen, M. Holmqvist, M. F. Yen, S. Abdsalah, B. Epstein, E. Frodis, E. Ljungberg, C. Hedborg-Melander, A. Sundbom, M. Tholin, M. Wiege, A. Akerlund, H. M. Wu, T. S. Tung, Y. H. Chiu, C. P. Chiu, C. C. Huang, R. A. Smith, M. Rosen, M. Stenbeck, and L. Holmberg, "The impact of organized mammography service screening on breast carcinoma mortality in seven swedish counties," *Cancer*, vol. 95, no. 3, pp. 458–69, 2002.

[3] D. B. Kopans, "Beyond randomized controlled trials: organized mammographic screening substantially reduces breast carcinoma mortality," *Cancer*, vol. 94, no. 2, pp. 580–1; author reply 581–3, 2002. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/11900247

[4] S. W. Duffy, L. Tabar, and R. A. Smith, "The mammographic screening trials: commentary on the recent work by Olsen and Gotzsche," *CA Cancer J Clin*, vol. 52, no. 2, pp. 68–71, 2002. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/11929006

[5] L. Tabar, B. Vitak, H. H. Chen, M. F. Yen, S. W. Duffy, and R. A. Smith, "Beyond randomized controlled trials: organized mammographic screening substantially reduces breast carcinoma mortality," *Cancer*, vol. 91, no. 9, pp. 1724–31, 2001. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/11335897

[6] A. N. Tosteson, D. G. Fryback, C. S. Hammond, L. G. Hanna, M. R. Grove, M. Brown, Q. Wang, K. Lindfors, and E. D. Pisano, "Consequences of false-positive screening mammograms," *JAMA internal medicine*, vol. 174, no. 6, pp. 954–961, 2014.

[7] D. B. Kopans, "An open letter to panels that are deciding guidelines for breast cancer screening," *Breast Cancer Res Treat*, vol. 151, no. 1, pp. 19–25, 2015. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/25868866

[8] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[9] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural computation*, vol. 1, no. 4, pp. 541–551, 1989.

[10] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[11] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *International conference on machine learning*, 2011, pp. 689–696.

[12] N. Srivastava and R. Salakhutdinov, "Multimodal learning with deep boltzmann machines," in *Advances in neural information processing systems*, 2012, pp. 2222–2230.

[13] W. Wang, R. Arora, K. Livescu, and J. Bilmes, "On deep multi-view representation learning," in *International Conference on Machine Learning*, 2015, pp. 1083–1092.

[14] H. Su, S. Maji, E. Kalogerakis, and E. Learned-Miller, "Multi-view convolutional neural networks for 3d shape recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 945–953.

[15] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error-propagation," in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 1*. MIT Press, Cambridge, MA, 1986, vol. 1, no. 6088, pp. 318–362.

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[17] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting." *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[19] M. Lin, Q. Chen, and S. Yan, "Network in network," in *International Conference on Learning Representations*, 2013.

[20] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representation*, 2015.

[21] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9.

[22] I. Domingues and J. S. Cardoso, "Mass detection on mammogram images: a first assessment of deep learning techniques," 2013.

[23] N. Dhungel, G. Carneiro, and A. P. Bradley, "Automated mass detection in mammograms using cascaded deep learning and random forests," in *International Conference on Digital Image Computing: Techniques and Applications*. IEEE, 2015, pp. 1–8.

[24] M. G. Ertosun and D. L. Rubin, "Probabilistic visual search for masses within mammography images using deep learning," in *IEEE International Conference on Bioinformatics and Biomedicine*. IEEE, 2015, pp. 1310–1315.

[25] B. Q. Huynh, H. Li, and M. L. Giger, "Digital mammographic tumor classification using transfer learning from deep convolutional neural networks," *Journal of Medical Imaging*, vol. 3, no. 3, pp. 034 501–034 501, 2016.

[26] D. Lévy and A. Jain, "Breast mass classification from mammograms using deep convolutional neural networks," *arXiv:1612.00542*, 2016.

[27] J. Arevalo, F. A. González, R. Ramos-Pollán, J. L. Oliveira, and M. A. G. Lopez, "Representation learning for mammography mass lesion classification with convolutional neural networks," *Computer methods and programs in biomedicine*, vol. 127, pp. 248–257, 2016.

[28] J.-J. Mordang, T. Janssen, A. Bria, T. Kooi, A. Gubern-Mérida, and N. Karssemeijer, "Automatic microcalcification detection in multi-vendor mammography using convolutional neural networks," in *International Workshop on Digital Mammography*. Springer, 2016, pp. 35–42.

[29] T. Kooi, G. Litjens, B. van Ginneken, A. Gubern-Mérida, C. I. Sánchez, R. Mann, A. den Heeten, and N. Karssemeijer, "Large scale deep learning for computer aided detection of mammographic lesions," *Medical image analysis*, vol. 35, pp. 303–312, 2017.

[30] A. S. Becker, M. Marcon, S. Ghafoor, M. C. Wurnig, T. Frauenfelder, and A. Boss, "Deep learning in mammography: Diagnostic accuracy of a multipurpose image analysis software in the detection of breast cancer." *Investigative Radiology*, 2017.

[31] A. Akselrod-Ballin, L. Karlinsky, S. Alpert, S. Hasoul, R. Ben-Ari, and E. Barkan, "A region based convolutional network for tumor detection and classification in breast mammography," in *International Workshop on Large-Scale Annotation of Biomedical Data and Expert Label Synthesis*. Springer, 2016, pp. 197–205.

[32] W. Zhu, Q. Lou, Y. S. Vang, and X. Xie, "Deep multi-instance networks with sparse label assignment for whole mammogram classification," *arXiv:1612.05968*, 2016.

[33] G. Carneiro, J. Nascimento, and A. P. Bradley, "Unregistered multiview mammogram analysis with pre-trained deep learning models," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 652–660.

[34] I. C. Moreira, I. Amaral, I. Domingues, A. Cardoso, M. J. Cardoso, and J. S. Cardoso, "Inbreast: toward a full-field digital mammographic database," *Academic radiology*, vol. 19, no. 2, pp. 236–248, 2012.

[35] K. Bowyer, D. Kopans, W. Kegelmeyer, R. Moore, M. Sallam, K. Chang, and K. Woods, "The digital database for screening mammography," in *Third international workshop on digital mammography*, vol. 58, 1996, p. 27.

[36] M. Heath, K. Bowyer, D. Kopans, P. Kegelmeyer Jr, R. Moore, K. Chang, and S. Munishkumaran, "Current status of the digital database for screening mammography," in *Digital mammography*. Springer, 1998, pp. 457–460.

[37] J. Wang, X. Yang, H. Cai, W. Tan, C. Jin, and L. Li, "Discrimination of breast cancer with microcalcifications on mammography by deep learning," *Scientific reports*, vol. 6, p. 27327, 2016.

[38] A. J. Bekker, H. Greenspan, and J. Goldberger, "A multi-view deep learning architecture for classification of breast microcalcifications," in *IEEE International Symposium on Biomedical Imaging*, 2016, pp. 726–730.

[39] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks." in *International Conference on Artificial Intelligence and Statistics*, 2010.

[40] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representation*, 2015.

[41] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European Conference on Computer Vision*. Springer, 2014, pp. 818–833.

[42] J. Yosinski, J. Clune, A. Nguyen, T. Fuchs, and H. Lipson, "Understanding neural networks through deep visualization," *arXiv:1506.06579*, 2015.

[43] J. S. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl, "Algorithms for hyper-parameter optimization," in *Advances in Neural Information Processing Systems 24*, 2011, pp. 2546–2554.

[44] J. Snoek, O. Rippel, K. Swersky, R. Kiros, N. Satish, N. Sundaram, M. M. A. Patwary, M. Prabhat, and R. P. Adams, "Scalable bayesian optimization using deep neural networks," in *International Conference on Machine Learning*, 2015.