Unsupervised deep learning applied to breast density segmentation and mammographic risk scoring

Michiel Kallenberg, Kersten Petersen, Mads Nielsen, Andrew Y. Ng, Pengfei Diao, Christian Igel, Celine M. Vachon, Katharina Holland, Rikke Rass Winkel, Nico Karssemeijer, and Martin Lillholm

***Abstract*—Mammographic risk scoring has commonly been automated by extracting a set of handcrafted features from mammograms, and relating the responses directly or indirectly to breast cancer risk. We present a method that learns a feature hierarchy from unlabeled data. When the learned features are used as the input to a simple classifier, two different tasks can be addressed: i) breast density segmentation, and ii) scoring of mammographic texture. The proposed model learns features at multiple scales. To control the models capacity a novel sparsity regularizer is introduced that incorporates both lifetime and population sparsity. We evaluated our method on three different clinical datasets. Our state-of-the-art results show that the learned breast density scores have a very strong positive relationship with manual ones, and that the learned texture scores are predictive of breast cancer. The model is easy to apply and generalizes to many other segmentation and scoring problems.**

***Index Terms*—Unsupervised feature learning, deep learning, breast cancer, mammograms, prognosis, risk factor, segmentation**

I. Introduction

BREAST cancer is the most frequently diagnosed cancer among women, worldwide [1]. In 2012, $464,000$ new cases (13.5% of all cancers) were diagnosed in Europe and $131,000$ died from the disease [2]. Breast cancer mortality can be reduced by identifying high risk patients early and treating them adequately [3]. One of the strongest known risk factors for breast cancer after gender, age, gene mutations, and family history is the relative amount of radiodense tissue in the breast, expressed as mammographic density (MD). According to several studies, women with high MD have a two to six-fold increased breast cancer risk compared to women with low MD [4], [5]. Further, breast density is modifiable and density changes relate to breast cancer risk. Tamoxifen, for example, reduces breast density and decreases the risk, whereas hormone replacement therapy causes the opposite [6].

Many MD scores have been proposed, ranging from manual categorical (e.g. BI-RADS) to automated continuous scores. In early years, radiologists characterized the mammographic appearance by a set of intuitive, but loosely defined breast tissue patterns that were shown to relate to the risk of breast cancer [7], [8]. The current gold standard are semi-automated continuous scores, as obtained by Cumulus-like thresholding [9]. In Cumulus, the radiologist sets an intensity threshold to separate radiodense (white appearing) from fatty (dark appearing) tissue. The computer then measures the proportion of dense to total breast area, known as percentage mammographic density (PMD). However, user-assisted thresholding is subjective and time-consuming, and hence not suited for large epidemiological studies. There has been a trend towards fully automating PMD scoring [10], [11], [12], [13], [14], [15], but most of these approaches rely on handcrafted features with several parameters that need to be controlled. Generalizing these methods beyond the reported datasets could be challenging.

Finding features that capture the relevant information in the mammogram is a difficult task. This becomes even more apparent when looking at work on mammographic texture (MT) scoring. MT scoring methods aim to find breast tissue patterns (or textures) that are predictive of breast cancer [16], [17], [18], [19], [20], [21], [22]. Intuitively, their goal is to characterize breast heterogeneity instead of breast density. MT scoring is even harder than MD scoring, since the label of interest (healthy vs. diseased) is defined per image and not per pixel (e.g. fatty vs. dense). Previous work on MT scoring has focused on manually designing and selecting features, similar to automatic MD scoring methods [17], [18], [19], [20]. However, these studies reach different conclusions on which texture features discriminate best. Furthermore, it is unclear if the published methods generalize to multiple datasets.

The goal of this paper is to present a method that automatically learns features for images, which in our case are mammograms. The model is called a *convolutional sparse autoencoder* (CSAE), as its core consists of a sparse autoencoder within a convolutional architecture. The method extends previous work on CSAEs [23], [24] to the problem of pixel-wise labeling and to large images (instead of small patches). The proposed CSAE is generic, easy to apply, and requires barely any prior knowledge about the problem. The main idea of the model is to learn a deep hierarchy of increasingly more abstract features from unlabeled data. Once the features have been learned, a classifier is trained to map the features to the labels of interest.

We evaluate the method on two breast-cancer tasks that have

Michiel Kallenberg and Mads Nielsen are with the University of Copenhagen, DK-2100 Copenhagen OE, Denmark, and also with Biomediq A/S, DK-2100 Copenhagen OE, Denmark.

Kersten Petersen, Pengfei Diao and Christian Igel are with the University of Copenhagen, DK-2100 Copenhagen OE, Denmark.

Andrew Y. Ng is with Stanford University, CA 94305, United States

Celine M. Vachon is with Mayo Clinic Hospital, AZ 85054, United States

Nico Karssemeijer and Katharina Holland are with Radboud University Nijmegen Medical Centre, 6525 GA Nijmegen, Netherlands

Rikke Rass Winkel is with University Hospital of Copenhagen, DK-2100 Copenhagen OE, Denmark

Martin Lillholm is with Biomediq A/S, DK-2100 Copenhagen OE, Denmark.

Michiel Kallenberg and Kersten Petersen contributed equally to this work.

previously been addressed in very different ways: The first task is the automated segmentation of breast density (MD). The second task is to characterize mammographic textural (MT) patterns with the goal of predicting whether a woman will develop breast cancer.

As in our previous work on multiscale denoising autoencoders [25], [26], we analyze features at multiple scales. On top of that, the CSAE employs a convolutional architecture that models the topology of images, and integrates a novel sparsity term to control the model capacity. We continue with a literature review for each of the two concerned tasks and summarize related work on feature learning.

A. Mammographic Density Scoring (MD)

Various approaches have been suggested to automate percentage mammographic density (PMD), which is widely considered as the gold standard in mammographic density scoring. A recent overview of methods can be found in He et al.[27]. A first class of methods takes the global image appearance into account. Sivaramakrishna et al. [28] mimicked PMD by measuring Kittler's optimal threshold, whereas Torrent et al. [29] determined the threshold based on excess entropy. Ferrari et al. [30] fitted a Gaussian Mixture Model to regions of different density. Keller et al. [15] utilized adaptive multiclass fuzzy c-means clustering on the gray-level intensity followed by support vector machine classification.

None of the aforementioned methods takes neighborhood information into account. To capture structural information, several authors assessed breast density using texture features from the computer vision literature. An approach that integrates many of these features with location, intensity, and global contextual information has been proposed by Kallenberg et al. [10]. The approach achieves state-of-the-art performance, but introduces a plethora of parameters that need to be controlled. To overcome this problem, we have recently proposed a feature learning method called multiscale denoising autoencoder [25],[26]. The method is more generic, yet achieves comparable results in automating MD.

Instead of assessing PMD in the breast area, it has also been suggested to estimate PMD in the breast volume [31], [32]. Highnam and Brady [31] suggested the standard mammographic form, a model of the imaging process, to automate volumetric PMD.

In this paper, we use a similar framework as in [25], [26], but introduce a convolutional learning architecture that preserves the spatial layout of the image and regularizes the learning algorithm with a novel sparsity term.

B. Mammographic Texture Scoring (MT)

Mammographic texture (MT) scores consider structural information of breast tissue and can be grouped into *manual* and *automated* MT scores. Manual MT scores characterize breast tissue by a small number of intuitive, but rather imprecise patterns. Popular examples include the Wolfe patterns [7] or the Tabár score [8]. In contrast, existing automated MT scores select a set of generic statistical features and employ a statistical learning algorithm to separate healthy from diseased patients. Consequently, automated MT scores may consider textural patterns that are predictive, but weakly correlated with manual density patterns.

The literature contains various approaches for automated MT scores. Byng et al. [33], Huo et al. [34], and Heine et al. [20] estimated texture by computing histogram statistics, such as the central moments or the entropy of the histogram. Also features that capture spatial relationships among pixels have been considered, such as statistics of the gray-level co-occurrence matrix (GLCM) [17], [18], run-length measures [17], [18], Laws features [17], Fourier techniques [17], Wavelet features [17], [18], fractal dimension [33], [29], or lacunarity [29]. Manduca et al. [17], Häberle et al. [18] and Zheng et al. [22] summarized and combined most of the common heuristic texture features for breast cancer risk assessment. The approaches resemble each other with respect to the examined features. However, they differ in the evaluated dataset, feature selection schemes, classifiers, and the region of interest for computing the MT score. Manduca et al. found that a set of Fourier and Wavelet features at coarse scales performs best, whereas Häberle et al. concluded that certain GLCM and histogram features from fine and coarse scales are most predictive. Zheng et al. found that extracting features from multiple locations in the breast outperforms a single-ROI approach.

Nielsen et al. [19] investigated another method to determine the texture features. They selected a combination of multiscale 3-jet and 2D location features, employed a sequential forward selection using bootstrapping, and predicted pixel-wise labels which were afterwards averaged over the breast region.

In contrast to previous work, we do not handpick heuristic texture features, but instead aim to learn meaningful texture features directly from the unlabeled mammograms. The hope is that an uncommitted method is better suited to generalize to different datasets.

C. Feature Learning

A lot of research has been devoted to selecting and handcrafting features that encode the important factors of variation in the input data. However, it can be time-consuming and tedious to mathematically describe human intuition and domain-specific knowledge. Furthermore, human heuristics are not guaranteed to capture the salient information of the data, and features that perform well on a related computer vision problem may not transfer to the application at hand.

An increasing number of papers demonstrate that comparable or even better results are achieved by learning features directly from the data. Especially deep nonlinear models have been proven to generate descriptors that are extremely effective in object recognition and localization in natural images. A recent overview of feature learning with deep models is given in [35] and [36]. Inspired by the human brain, these architectures first learn simple concepts (or features) and then compose them to more complex ones in deeper layers. In addition, features share components from lower layers which allow them to compactly express the idiosyncrasies of the data and fight the curse of dimensionality [37]. Most of these models are trained by iteratively encoding features (forward propagation) and updating the learned weights to improve the optimization (backward propagation).

One approach is to jointly optimize the features of the deep model, in order to minimize the loss between the predictions of the top most layer and the target values. Traditional neural networks fall into this category, and also variants like convolutional neural networks (CNNs) by Lecun et al. [38], which are tailored towards images. Deep neural networks, such as CNNs, have been successfully applied to challenging image analysis problems, e.g., object recognition, scene parsing [39], cell segmentation [40], neural circuit segmentation [41], [42], analysis of images the breast [43], [44], [45], [46]. They were found to be faster and more expressive than other graphical models like Markov or Conditional Random Fields [47].

The features can also be learned in an unsupervised way, e.g. using Restricted Boltzmann Machines [48], [49] or autoencoders [23], [50], [51]. The features are typically learned in a greedy, layer-wise fashion, before a classifier is trained to predict the labels from the feature responses of the top most layer. The division into multiple optimization problems has several advantages. First, large amounts of unlabeled data can be exploited for training the features. Second, the features are learned faster and more stable, as each layer is optimized by a small encoder-decoder architecture instead of a complex deep network. And third, these deep models can incorporate transformations and classifiers that are optimized independently from the features.

In this paper, we employ a *sparse* autoencoder for learning the features in an unsupervised way. Previous work has suggested sparse autoencoders for object recognition from small image patches [23], [24], [52]. In contrast, we propose a feature learning method for images that exploits information at multiple scales and incorporates a different sparsity regularizer.

II. Method

We explain the overall approach consisting of three parts: generating input data, model representation, and parameter learning. The input data is composed of multiscale image patches that capture both detail and large contextual regions. The patches are processed by a multilayer convolutional architecture. The parameters of this representation are learned using a sparse autoencoder, which enhances the standard autoencoder with a novel sparsity regularizer.

A. Overall Approach

Assume we are given a set of training images with associated label masks and our goal is to predict the label mask for an unseen image. It would be computationally prohibitive to map entire images to label masks. Downsampling the image is also infeasible, as many structures of interest occur at a fine scale. However, we can learn a compact representation for local neighbors (or *patches*) from the image.

Let us represent the labels in a 1-of-C coding scheme. Then formally, we aim to map a multi-channel image patch $x \in \mathcal{X} = \mathbb{R}^{c \times m \times m}$ of size $m \times m$ with c channels to a label posterior patch $y \in \mathcal{Y} = \mathbb{R}^{C \times M \times M}$ of size $M \times M$ with one channel per label, where we assume quadratic input sizes for ease of notation. The image and label posterior patch are centered at the same location, but can have different sizes. The channels of the image patch may include color channels, preprocessed image patches, or feature responses.

For training our model, n labeled training examples $\mathcal{D} = \{(x^{[i]}, y^{[i]})\}_{i=1}^{n}$ are extracted at randomly chosen locations across the set of training images. Given the training data $\mathcal{D}$, our model learns a hypothesis function $h : \mathcal{X} \mapsto \mathcal{Y}$ which is parameterized by θ.

In this paper, the hypothesis function h is defined as a latent variable model that consists of multiple layers. Instead of mapping x to y directly, we learn a series of increasingly more abstract *feature representations*[1] $z^{(l)}$ for layers $l \in 1, \dots, L$, where $z^{(1)} = x$ and $z^{(L)} \in \mathcal{Y}$. The feature representations are gained by encoding the input through a cascade of transformations, of which some are trainable. We learn the parameters of these transformations in a greedy layer-wise fashion without using the labels. While an individual layer is not deep, the stacked architecture is (e.g, the second layer receives as input the output from the first layer). Thus, the individual unsupervised training of ("shallow") layers results in an unsupervised deep learning procedure.

Three steps are necessary to move from one feature representation, $z^{(l)}$, to the next one, $z^{(l+1)}$:

1) Extract sub-patches (called *local receptive fields*) from random locations in $z^{(l)}$ and optionally preprocess them.
2) *Feature learning:* Learn transformation parameters (or features) by *autoencoding* the local receptive fields.
3) *Feature encoding:* Transform all local receptive fields in $z^{(l)}$ using the learned features from step 2. The result of the transformation is referred to as the *feature representation* $z^{(l+1)}$.

A *classifier* maps the last feature representation into label space $\mathcal{Y}$. An unseen image is tested by applying the trained hypothesis function $h_\theta(x)$ to all possible patches in a sliding window approach. Thus, every patch within the tested image is sent through the trained encoders and classifier to create a prediction. If the size of the predicted output region is bigger than a single pixel, i.e., $M > 1$, predictions at neighboring image locations might overlap with each other. These predictions can be fused by computing the average probability per class.

An overview of the pipeline is shown in Fig. 1. Our architecture consists of four hidden layers: a convolutional layer, a maximum pooling layer, and two further convolutional layers. We chose one pooling layer to be invariant towards small distortions, but sensitive to fine-scaled structures. The specifics will be presented in the following sections.

B. Multiscale Input Data

We capture long range interactions in the mammograms by extracting input examples x from multiple scales. As introduced

[1] We use the terms *weights* and *features* interchangeably to refer to the parameters of a representation transformation. The output of this transformation are called *activations* or *feature representation*. Within a convolutional architecture, the activations will be spatially arranged as *feature maps* (see Section II).

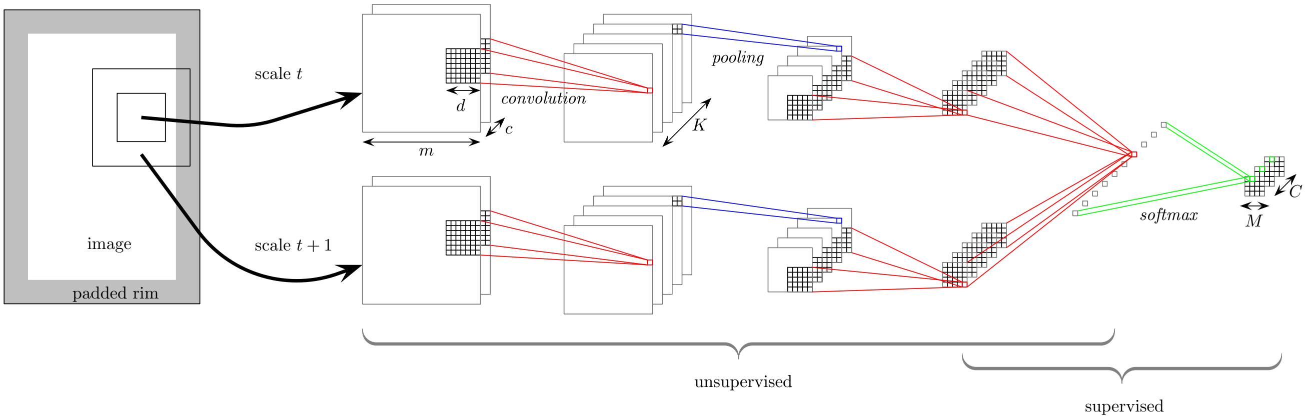


Fig. 1: Deep convolutional architecture consisting of convolutional, pooling and a softmax layer(s). Input patches are extracted from multiple scales of an image. The pixel spacing of the patches is adjusted such that the feature maps at different scale levels are equally sized. Each scale level of the CSAE model is processed in isolation before all activations are integrated in the second last layer. The convolutional layers in the unsupervised parts are trained as autoencoders; In the supervised part the (pretrained) weights and bias terms are fine-tuned using softmax regression (see text for details).

in our previous work [25], [26], a given mammogram I is embedded into a Gaussian scale space $I(u;\sigma_t) = [I * G_{\sigma_t}](u)$. Here the $*$ operator denotes convolution. Multi-scale mammographic analysis is realized using the well established discrete scale space theory (see, e.g., [53]); specifically we use a Fourier implementation where the Gaussian kernel is discretized in the Fourier domain and spatial convolution obtained through multiplication (in the Fourier domain) with the discrete Fourier transform of the mammogram [54]. The parameter $s \in \mathbb{R}^2$ denotes the position (or site) and σ_t determines the standard deviation of the Gaussian at the tth scale. More specifically, the standard deviation

$$\sigma_t = \sqrt{\sum_{i=0}^{t-1} \delta^{2i}} \tag{1}$$

is given as the square root of the summed Gaussian variances from the first t scale levels of the Gaussian pyramid. In this paper, we chose downsampling factor $\delta = 2$.

An input example x_t at location u from scale t is constructed by sampling a patch with pixel distance (or stride) δ^{t-1} around location u in the Gaussian scale space. For example, an input patch at scale level $t = 1$ is a coherent $m \times m$ region, whereas the patch at scale $t = 4$ considers only every eighth pixel around u from a heavily smoothed mammogram.

The underlying representation of our model, a convolutional architecture, processes inputs from multiple scales (Fig. 1). For computational reasons, features are first learned for each scale in isolation, before they are merged in deeper layers.

C. Sparse Autoencoder

It would be possible to learn the weights (or features) using forward and backward propagation through the entire architecture [38]. However, as argued in our review of feature learning, we aim to learn features in an unsupervised way using autoencoders. We propose a variant of the autoencoder that enables to learn a *sparse overcomplete* representation. A feature representation is called *overcomplete* if it is larger than the input. *Sparsity* forces most of the entries to be zero, leaving only a small number of non-zero entries to represent the input signal. Thus, in the case of extreme sparsity, each input example would be encoded by a single hidden unit, the one whose input weights (or feature) are the most similar to the input example.

Sparse overcomplete representations provide simple interpretations, are cost-efficient, and robust to noise. They are suited to disentangle the underlying factors of variation because each input example needs to be represented by the combination of a few (specialized) features.

In previous work, feature representations have been made sparse by limiting the number of active (non-zero) units per example (*population sparsity*) or by limiting the number of examples for which a specific unit is active (*lifetime sparsity*). Population sparsity underlies methods like sparse coding [55], or K-means, where each cluster centroid can be interpreted as a feature and each example is encoded by the most similar centroid. Lifetime-sparsity is incorporated in the sparsifying logistic by Ranzato et al. [23] or the sparse RBM by Lee et al. [56], where the average activation per unit is supposed to equal a user-specified sparsity threshold.

In this paper, we formulate a sparsity regularizer that incorporates both population sparsity and lifetime sparsity. While population sparsity enforces a compact encoding per example, lifetime sparsity leads to example-specific features. Our proposed sparsity prior can be combined with any activation function including the rectified linear function, which was shown to produce better features than the sigmoid or the hyperbolic tangent in [57]. The formalization of the sparse autoencoder is given in the appendix.

D. Experiments and Datasets

We evaluated the performance of the CSAE for two different tasks (MD, MT) on three different mammographic datasets. For each task we first segmented mammograms into background, pectoral muscle, and breast tissue region. The breast tissue region was then used as a region of interest for the mammographic scoring tasks (MD and MT). We continue with a description of the datasets, the parameter settings, and the results for each of the two tasks.

1) Density Dataset: From the Dutch breast cancer screening program we collected 493 mammograms of healthy women. Mean age of the women was 60.25 ± 7.83 years. The images were recorded between 2003 and 2012 on a Hologic Selenia FFDM system, using standard clinical settings. We used the raw image data. The set contained a mixture of mediolateral oblique (MLO) and craniocaudal (CC) views from the left and right breast. For each woman however only one view was available.

A trained radiologist annotated the skin-air boundary and the pectoral muscle by a polygon tool. In a second step, the breast tissue area was delineated by cropping superfluous tissue folds below and above the breast area. The radiologist estimated percent density using a Cumulus like approach.

2) Texture Dataset: The texture dataset comprises 668 mediolateral mammograms from the Mayo mammography Health Study (MMHS) cohort at the Mayo Clinic in Rochester, Minnesota. The purpose of the MMHS study was to examine the association of breast density with breast cancer [58]. The chosen subset included 226 cases and 442 controls that were matched on age and time from earliest available mammogram to study enrollment/diagnosis date. The images were recorded between October 2003 and September 2006, between 6 months and 15 years prior to the detection of the cancer. The mean age was 55.2 ± 10.5 years.

All mammograms were digitized with an Array 2905 laser digitizer (Array Corporation, the Netherlands) that provided a pixel spacing of 50 microns on a 12-bit gray scale. A trained observer annotated the skin-air boundary and the pectoral muscle by a polygon tool.

3) Dutch Breast Cancer Screening Dataset: From the Dutch breast cancer screening program we collected 394 cancers, and 1182 healthy controls. Controls were matched on age and acquisition date. The images were recorded between 2003 and 2012 on a Hologic Selenia FFDM system, using standard clinical settings. For each woman MLO views from both the right and left breast were available. However, to exclude signs of cancerous tissue, we took the contralateral mammograms for our analyses on breast cancer risk prediction. We used the raw image data. Mean age of the women was 60.6 ± 7.70 years. The images were segmented into the breast area, pectoral muscle and background using automated software (Volpara, Matakina Technology Limited, New Zealand).

E. Parameter Settings and Model Selection

If not stated otherwise, the same parameter settings have been applied to each task and each dataset.

1) Patch Creation: Before extracting the patches, the mammograms were resized to an image resolution of roughly 50 pixels per mm. The model was trained on $n = 48,000$ patches. The patch size in terms of number of pixels was restricted to 24x24 in order to keep the number of trainable weights and bias terms limited. The training patches were sampled across the whole dataset as follows: For density scoring 10% of the patches were sampled from the background and the pectoral muscle, 45% from the fatty breast tissue, and 45% from the dense breast tissue. For texture scoring 50% of the patches were sampled from the breast tissue of controls, and 50% from the breast tissue of cancer cases. In pilot experiments we experimented with different breast tissue masks to sample patches from. Best results were obtained if we restricted the sampling of the patches to the inner breast zone, which is the breast area that is fully compressed during image acquisition, and in which the fibroglandular tissue is most prominent. For both tasks $M = 1$ was chosen. We set scales t to 1 to 4 for both density and texture scoring. The smallest patch was thus 4.8mm x 4.8mm, whereas the biggest patch was 3.7cm x 3.7cm. As such several structures of interest could be captured in different detail. On a validation set we experimented with different setups of the input channels. Best results were obtained by having one input channel consisting of the unprocessed image.

2) Convolutional Architecture: For each tasks the number of feature map were set to $K = \{50, (50), 50, 100\}$; the associated kernel sizes were fixed to $\{7, 2, 5, 5\}$. These values were motivated from previous work on convolutional architectures [59].

3) Sparse Autoencoder: To learn the weights of the convolutional layers, a sparse autoencoder was trained on $N = 48,000$ extracted local receptive fields from the activations of the previous layer. For the first layer each local receptive field was preprocessed by removing its DC components. The sparsity parameter was set to $\rho = 0.01$ and the weighting term of the sparsity regularizer to $\lambda = 1$. We applied the backpropagation algorithm to compute the gradient of the objective function in (6). The parameters were optimized with L-BFGS using 25 mini-batches of size $2,000$. Each mini-batch was used for 20 iterations, such that the entire optimization ran for 500 iterations. In pilot experiments we determined the settings of the hyperparameters. In these pilot experiments we put most emphasis on the sparsity regularizer λ and the length of the training for both the unsupervised and the supervised part of our network. We found that the performance was robust for a broad range of values of the mentioned parameters.

4) Classifier: We trained a two layer neural network, consisting of a pretrained convolutional layer (i.e., layer L-1) and multinomial logistic regression (or softmax classifier) layer. That is, that the weights and bias terms of the pretrained convolutional layer (i.e., layer L-1) are fine-tuned with a supervised signal. For MD scoring we utilized three class labels: (i) pectoral muscle and background, (ii) fatty tissue, and (iii) dense tissue. For MT scoring we had two class labels: (i) cancer, and (ii) control. The optimization was performed for 500 iterations using L-BFGS on the n encoded patches. Unless stated otherwise for each task and dataset results were obtained by performing 5-fold cross-validation by image to estimate the generalization ability of our machinery.

III. Results

A. Mammographic Density Scoring

1) Density Dataset: The initial output of the MD scoring is a score that represents the posterior probability that a given pixel belongs to the dense tissue class. By thresholding the posteriors with threshold T_{dense} we obtain a segmentation of the dense tissue. Percent density (PMD) is then computed as the percentage of breast pixels that is segmented as dense. To speed up training we oversampled the dense class during training. As such our machinery tends to overestimate the density if we set the threshold T_{dense} to 0.50. By raising T_{dense} this effect is compensated for. Figure 2 shows the effect of T_{dense} on two performance measures, namely (i) the image-wise average of the Dice coefficient, defined as $2|A \cap B|/(|A| + |B|)$ between the automated segmentation A and the segmentation of the radiologist B, and (ii) the root mean squared error between the percent density (PMD) as measured by our machinery and the radiologist. Best results are obtained with T_{dense} in the interval 0.70-0.80. In the remainder of the paper results are therefore reported with T_{dense} set to 0.75. Table I summarizes the results on the density dataset. Reported are (i) the Pearson correlation coefficient (and 95% CI) between PMD as measured by our machinery and the radiologist, (ii-iii) the image-wise average (± standard deviation) of the Dice coefficient for both dense and fatty tissue, and (iv) the average percent density (± standard deviation). Figure 3 shows an example of a mammogram, the corresponding Cumulus-like segmentation and the segmentation obtained with the CSAE that incorporates the novel sparsity term.

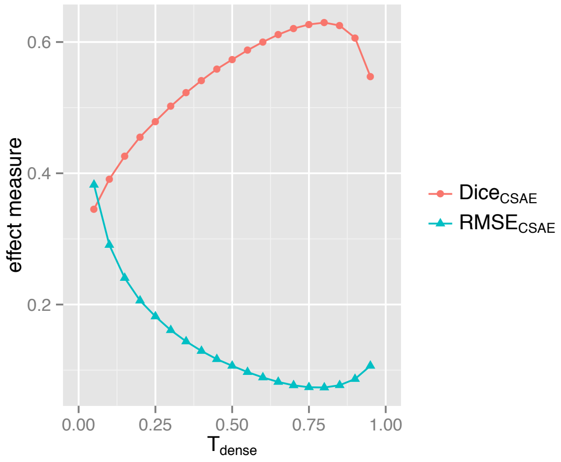
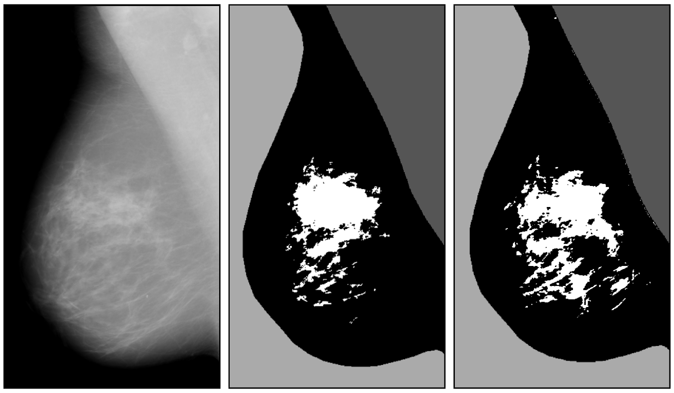


Fig. 2: Effect of varying the threshold on the posteriors T_{dense} on two performance measures of MD scoring, namely (i) the image-wise average of the Dice coefficient, and (ii) the root mean squared error between the percent density (PMD) as measured by our machinery and the radiologist.

2) Dutch Breast Cancer Screening Dataset: We used the networks that were trained on the density dataset to score PMD on all images of the Dutch Breast Cancer Screening Dataset.

TABLE I: Comparison of automated with radiologist's MD scores for the density dataset.

$R_{\text{PMD}_{\text{CSAE}}\text{-PMD}_{\text{Rad}}}$	0.85 (0.83-0.88)
$\text{Dice}_{\text{dense}}$	0.63 ± 0.19
Dice_{fat}	0.95 ± 0.05
PMD	0.16 ± 0.11



(a) (b) (c)

Fig. 3: Automated MD thresholding. Depicted are (a) original image, (b) dense tissue according to expert Cumulus-like threshold, and (c) dense tissue according to CSAE .

Subsequently we assessed how well our estimation of PMD is able to discriminate between cancers and controls. Table II presents (i) left-right correlation for the automated PMD scores (ii-iii) mean and standard deviation of the PMD scores for cancers and controls, and (iv) the area under the ROC curve (AUC) for separating between cancers and controls.

TABLE II: Statistics of MD scores on the Dutch Breast Cancer Screening dataset.

$R_{\text{PMD}_{\text{left}}\text{-PMD}_{\text{right}}}$	0.93 (0.92-0.94)
PMD_{Case} (n=394)	0.19 ± 0.11
$\text{PMD}_{\text{Control}}$ (n=1182)	0.15 ± 0.11
AUC_{PMD}	0.59 (0.56-0.62)

B. Mammographic Texture Scoring

1) Texture dataset: The initial output of the MT scoring is a score that represents the posterior probability that a given pixel belongs to the cancer class. To obtain one MT score per image we averaged the posteriors of 500 patches randomly sampled from the breast area. We have evaluated the MT scoring performance on the texture dataset (see Table III). Our model improved on two state-of-the-art methods in MT scoring: (i) the KNN method by Nielsen et al. [19] using multiscale local jet features [60], which so far had reported the best results on the texture dataset (results were communicated); (ii) a softmax classifier on static histogram features inspired by the method of Häberle et al. [18]. A precise reimplementation of the original method by Häberle et al. was not possible, since we could not get access to important hyperparameters like the orientation of the chosen features. The static histogram features represent 16 of the 45 final selected features, but accounted for 15 of the 18 highest coefficients in their final softmax classifier.

We also checked the robustness of our results with respect to different randomizer seed points. We found that the CSAE model was able to produce similar scores in different runs. The AUC varied less than 0.01 across multiple runs.

TABLE III: AUC values for separating between cancers and controls for various automated MT scores on the texture dataset.

Method	AUC
Static histograms [18]	0.56 (0.51-0.61)
Multiscale local jet [19]	0.60 (N/A)
CSAE	0.61 (0.57-0.66)

2) Dutch Breast Cancer Screening Dataset: Table IV presents performance indicators for our MT scoring on the Dutch Breast Cancer Screening dataset. Shown are i) left-right correlation of the MT scores ii) the area under the ROC curve (AUC) for separating between cancers and controls.

TABLE IV: Statistics of MT scores on the Dutch Breast Cancer Screening dataset.

$R_{\text{MT}_{\text{left}}\text{-MT}_{\text{right}}}$	0.91 (0.90-0.92)
AUC_{MT}	0.57 (0.54-0.61)

IV. Discussion

We have presented an unsupervised feature learning method for breast density segmentation and automatic texture scoring. The model learns features across multiple scales. Once the features are learned, they are fed to a simple classifier that is specific to the task of interest. After adapting a small set of hyperparameters (feature scales, output size, and label classes), the CSAE model achieved state-of-the-art results on each of the tasks.

The results suggest that the proposed method was able to learn useful features for each of the considered applications. The automated PMD scores have a very strong positive relationship with the manual Cumulus scores ($R = 0.85$) and are competitive with reported correlation coefficients from the literature, e.g., 0.63 [61], 0.70 [12], 0.85 [15], 0.88 [14] and 0.91 [10]. We also evaluated how well the automated PMD scores separated out cases from controls. We found that the automated PMD scores yielded an AUC of 0.59, which is competitive to reported AUCs in the literature on similar populations (e.g. 0.57 [61], 0.59 [14], and 0.60 [62]). Thus, our automatic MD scoring method could be an alternative to subjective and expensive manual MD scoring.

The automated MT scores separated cancers and controls better than two state-of-the-art MT scoring methods. In the texture dataset the CSAE model improved on the KNN method by Nielsen et al. [19] and a simplified version of the model of Häberle et al. [18]. The full model of Häberle et al. could not be tested, as necessary parameter settings were missing.

Based on our results we conclude that useful discriminative features can be attained by "letting the data speak" instead of modeling prior assumptions.

We proposed a novel sparsity regularizer that incorporates both population sparsity and lifetime sparsity. We compared the performance of the machinery with the novel sparsity term with a control setup that used an alternative sparsity term [56], which measured the KL-divergence between the mean activation and the desired activation. For each experiment the novel sparsity term performed at least equally well as the control setup.

The stack of convolutional (sparse) autoencoders (CSAE) presented in this work forms a convolutional neural network (CNN). The major difference between a CSAE and a classic CNN is the usage of unsupervised pre-training. In our previous work [25] we found that unsupervised pre-training with autoencoders led to an increase in performance on similar tasks as presented here. This is in line with several works (e.g., [24], [63], [64], [65]) that demonstrated the merits of employing unsupervised pre-training with autoencoders in convolutional architectures.

We have focused on presenting a principled and generic framework for learning image features. The MT features were learned on image patches and mapped to individual locations in the image. In a second step, the classifier predictions were merged to assign a disease label for the mammogram. However, the labels in the texture scoring task are provided per mammogram. We assumed that texture changes are systemic and occur at many locations in the tissue. One may also hypothesize the opposite. Texture changes could be restricted to the vicinity of future cancers. We plan to extend the framework to learn from multiple instances. The idea would be to train a classifier that maps the feature responses from multiple locations to one label. This is a difficult task and probably requires many more disease labels than considered in this paper. However, with the advent of large screening datasets, it may become possible to learn a relationship from images to labels, and investigate the locality of texture changes.

The model could be easily adjusted to support 3D data. Features could be learned for different mammographic projections (e.g., craniocaudal views) or images from complementary modalities (e.g., ultrasound, magnetic resonance imaging, tomosynthesis, or computed tomography). There are several applications for automatically derived MD and MT scores. As part of a risk prediction model, they stimulate research on breast cancer epidemiology. For instance, large databases of historical mammograms could be scored to investigate change of breast cancer risk. Moreover, mammographic risk scores may affect decision making for the individual patient, e.g., the selection of screening interval, imaging modalities, or treatment options. Thus, they could help organize mammographic screening programs more efficiently and effectively, which may ultimately lead to a reduction in breast cancer mortality.

Appendix

In the unsupervised part of our machinery features are learned using autoencoders. We propose a variant of the autoencoder that enables to learn a *sparse overcomplete* representation. We introduce a novel sparsity regularizer that combines population sparsity and lifetime sparsity. We summarize the idea of the standard autoencoder (Fig. 4), before introducing an autoencoder that exploits sparsity.

A. Autoencoder

Consider learning the weights $w_j \in \mathbb{R}^{c \times d \times d}$ in for $j = 1, \ldots, K$, where we omit the layer index for brevity. We rewrite the K 3D weight arrays as a weight matrix $W \in \mathbb{R}^{K \times cd^2}$, where the jth row corresponds to w_j. Similarly, the bias vector $b \in \mathbb{R}^K$ concatenates the K bias terms b_j. Assume further that we have sampled one local receptive field at a random location per input feature map example $z^{[i]} \in \mathbb{R}^{c \times m \times m}$ with $i = 1, \ldots, n$. The local receptive fields have a size of $c \times d \times d$, but are arranged as vectors $r^{[i]} \in \mathbb{R}^{cd^2}$, where $i = 1, \ldots, n$ and $d \leq m$. Then, we can learn W and b in an unsupervised way by *autoencoding* the local receptive fields.

The autoencoder reconstructs an input $r \in \mathbb{R}^{cd^2}$ by a composition $f(g(r))$ of an encoder $g(\cdot)$ and a decoder $f(\cdot)$. The *encoder*

$$a \equiv g(r) = \phi(Wr + b) \tag{2}$$

connects the input layer with the hidden layer and uses the activation function $\phi(\cdot)$, which is commonly one of the following: the sigmoid, the hyperbolic tangent, or the recently introduced rectified linear function $\phi(x) = \max(0, x)$ that is used in this paper due to its reported superior performance [57]. The *decoder*

$$f(a) = \psi(Va + \tilde{b}) \tag{3}$$

is an affine mapping between the hidden layer and the output layer. The activation function of the decoder $\psi(\cdot)$ is usually set to the identity function and the weight matrix $V = W^\top$ is defined as the transpose of the encoder weight matrix (i.e., we use tied weights [66]). The bias of the decoder $\tilde{b} \in \mathbb{R}^{cd^2}$ has the same dimension as the input. Tying the weights of the encoder and decoder encourages V and W to be at the same scale and orthogonal to each other [67]. It also decreases the number of trainable parameters and thereby improves the numerical stability of the algorithm. The specialized decoder is thus given by $f(a) = W^\top a + \tilde{b}$.

Let us denote the set of training examples as $\mathcal{D}_{\text{rec}} = \{r^{[i]}\}_{i=1}^N$ and the trainable parameters as $\theta_{\text{rec}} = \{W, b, \tilde{b}\}$. Then the objective function to be minimized is

$$\mathcal{J}_{\text{AE}}(\mathcal{D}_{\text{rec}}, \theta_{\text{rec}}) = \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}_{\text{rec}} \left[r^{[i]}, f(g(r^{[i]})) \right], \tag{4}$$

where the reconstruction error

$$\mathcal{L}_{\text{rec}} \left[r^{[i]}, f(g(r^{[i]})) \right] = \| r^{[i]} - f(g(r^{[i]})) \|^2 \tag{5}$$

is the squared loss. To avoid that the autoencoder learns the identity function, the hidden layer is constrained to be *undercomplete*, i.e., the number of hidden units is smaller than the number of input units ($K < cd^2$).

B. Sparse autoencoder

We define a sparse autoencoder that minimizes the objective function

$$\mathcal{J}_{\text{SAE}}(\mathcal{D}_{\text{rec}}, \theta_{\text{rec}}) = \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}_{\text{rec}} \left[r^{[i]}, f(g(r^{[i]})) \right] + \lambda \omega_{\text{sp}}(A) \tag{6}$$

using the novel sparsity term

$$\omega_{\text{sp}}(A) = \omega_{\text{psp}}(A) + \omega_{\text{lsp}}(A) \ . \tag{7}$$

This regularizer combines population sparsity $\omega_{\text{psp}}(A)$ and lifetime sparsity $\omega_{\text{lsp}}(A)$ with respect to the activation matrix $A \in \mathbb{R}^{K \times n}$, $A_{ji} = a_j^{[i]} = g(r_j^{[i]})$.

To define the population sparsity term, let us compute the average absolute activation for the jth activation unit (averaged across the n examples)

$$\begin{aligned} \hat{\rho}_j &= \frac{1}{n} \sum_{i=1}^{n} |A_{ji}| \\ &= n^{-1} \| A_{j\cdot} \|_1 \ , \end{aligned} \tag{8}$$

where $\|A_{j\cdot}\|_1$ is the L_1-norm of the jth row in A. We compare this unit-wise population sparsity to a pre-specified sparsity parameter ρ

$$\omega_{\text{psp}}(A) = \frac{1}{K} \sum_{j=1}^{K} \tau(\hat{\rho}_j; \rho)^2 \tag{9}$$

and average the squared thresholded difference over the K units. Here, the threshold function

$$\tau(\hat{\rho}; \rho) = \max(\hat{\rho} - \rho, 0) \ . \tag{10}$$

penalizes sparsity values above ρ to avoid non-specific features. Values below ρ are not punished because selective features shall be permitted. A typical value for the sparsity level is $\rho = 0.01$ (see Section II-E).

Similarly, we specify the lifetime sparsity for the ith example as its average absolute activation averaged across the K activation units

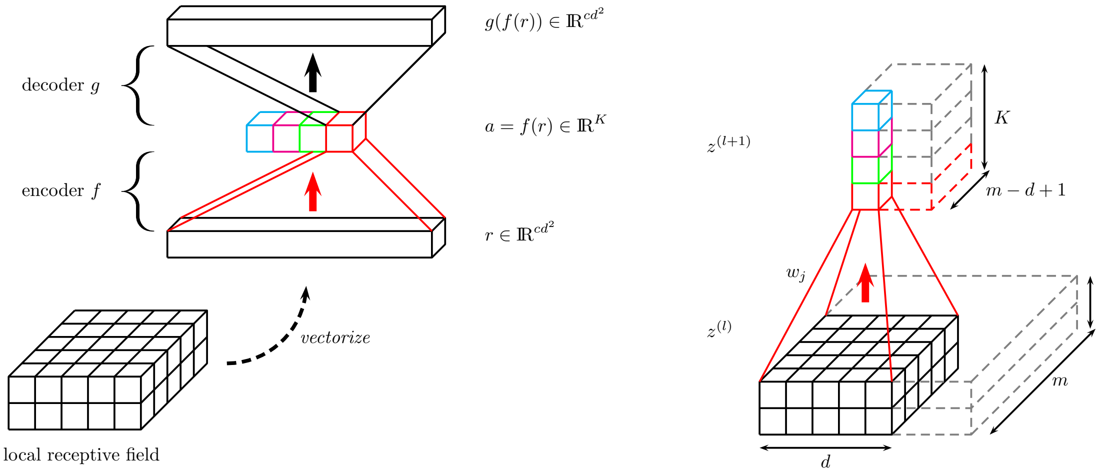
$$\begin{aligned} \hat{\rho}^{(i)} &= \frac{1}{K} \sum_{j=1}^{K} |A_{ji}| \\ &= K^{-1} \| A_{\cdot i} \|_1 \ , \end{aligned} \tag{11}$$

where $\|A_{\cdot i}\|_1$ is the L_1-norm of the ith column in A. The total lifetime sparsity is then given by

$$\omega_{\text{lsp}}(A) = \frac{1}{n} \sum_{i=1}^{n} \tau(\hat{\rho}^{[i]}; \rho)^2 \ . \tag{12}$$

Acknowledgments

This work was supported by the Innovation Fund Denmark under the grant 141-2013-6, the Danish National Advanced Technology Foundation under grant Personalized Breast Cancer Screening, the European Seventh Framework Programme FP7 under grant agreement no 306088, the European Commission through project AKMI (PCIG10-GA-2011-303655).



(a)

(b)

Fig. 4: (a) An autoencoder for learning the features of the convolutional layer. The input is vectorized and reconstructed by an encoder-decoder architecture. (b) Inference in a convolutional layer using a 3D convolution. The encoded units correspond to the highlighted units in output $z^{(l+1)}$ of the convolutional layer. The weights w_j between input feature maps $z^{(l)}$ and the jth output feature map are marked in red and initialized with the learned weights from the autoencoder. We refer to the text for details.

References

[1] A. Jemal, F. Bray, M. M. Center, J. Ferlay, E. Ward, and D. Forman, “Global cancer statistics,” *CA Cancer J Clin*, vol. 61, pp. 69–90, 2011. [Online]. Available: http://dx.doi.org/10.3322/caac.20107

[2] J. Ferlay, E. Steliarova-Foucher, J. Lortet-Tieulent, S. Rosso, J. W. W. Coebergh, H. Comber, D. Forman, and F. Bray, “Cancer incidence and mortality patterns in europe: estimates for 40 countries in 2012,” *Eur J Cancer*, vol. 49, pp. 1374–1403, 2013. [Online]. Available: http://dx.doi.org/10.1016/j.ejca.2012.12.027

[3] I. T. Gram, E. Funkhouser, and L. Tabár, “The tabár classification of mammographic parenchymal patterns,” *Eur J Radiol*, vol. 24, pp. 131–136, 1997.

[4] V. McCormack and I. dos Santos Silva, “Breast density and parenchymal patterns as markers of breast cancer risk: a meta-analysis,” *Cancer Epidemiol Biomarkers Prev*, vol. 15, pp. 1159–1169, 2006.

[5] N. F. Boyd, L. J. Martin, M. Bronskill, M. J. Yaffe, N. Duric, and S. Minkin, “Breast tissue composition and susceptibility to breast cancer,” *J Natl Cancer Inst*, vol. 102, pp. 1224–1237, 2010. [Online]. Available: http://dx.doi.org/10.1093/jnci/djq239

[6] J. Cuzick, J. Warwick, E. Pinney, S. W. Duffy, S. Cawthorn, A. Howell, J. F. Forbes, and R. M. Warren, “Tamoxifen-induced reduction in mammographic density and breast cancer risk reduction: a nested case–control study,” *Journal of the National Cancer Institute*, vol. 103, no. 9, pp. 744–752, 2011.

[7] J. N. Wolfe, “Risk for breast cancer development determined by mammographic parenchymal pattern,” *Cancer*, vol. 37, pp. 2486–2492, 1976.

[8] L. Tabár, S. W. Duffy, B. Vitak, H.-H. Chen, and T. C. Prevost, “The natural history of breast carcinoma,” *Cancer*, vol. 86, no. 3, pp. 449–462, 1999.

[9] J. W. Byng, N. F. Boyd, E. Fishell, R. A. Jong, and M. J. Yaffe, “The quantitative analysis of mammographic densities,” *Phys Med Biol*, vol. 39, pp. 1629–1638, 1994.

[10] M. G. Kallenberg, M. Lokate, C. H. van Gils, and N. Karssemeijer, “Automatic breast density segmentation: an integration of different approaches,” *Phys Med Biol*, vol. 56, pp. 2715–2729, 2011.

[11] S. Petroudi, T. Kadir, and M. Brady, “Automatic classification of mammographic parenchymal patterns: A statistical approach,” in *Engineering in Medicine and Biology Society, 2003. Proceedings of the 25th Annual International Conference of the IEEE*, vol. 1. IEEE, 2003, pp. 798–801.

[12] J. J. Heine, M. J. Carston, C. G. Scott, K. R. Brandt, F.-F. Wu, V. S. Pankratz, T. A. Sellers, and C. M. Vachon, “An automated approach for estimation of breast density,” *Cancer Epidemiol Biomarkers Prev*, vol. 17, pp. 3090–3097, 2008.

[13] A. Oliver, X. Llado, R. Marti, J. Freixenet, and R. Zwiggelaar, “Classifying mammograms using texture information,” in *Medical Image Understanding and Analysis*, 2007, pp. 223–227.

[14] J. Li, L. Szekely, L. Eriksson, B. Heddson, A. Sundbom, K. Czene, P. Hall, and K. Humphreys, “High-throughput mammographic-density measurement: a tool for risk prediction of breast cancer,” *Breast Cancer Res*, vol. 14, p. R114, 2012.

[15] B. M. Keller, D. L. Nathan, Y. Wang, Y. Zheng, J. C. Gee, E. F. Conant, and D. Kontos, “Estimation of breast percent density in raw and processed full field digital mammography images via adaptive fuzzy c-means clustering and support vector machine segmentation,” *Med Phys*, vol. 39, no. 8, pp. 4903–4917, 2012.

[16] H. Li, M. L. Giger, O. I. Olopade, and M. R. Chinander, “Power spectral analysis of mammographic parenchymal patterns for breast cancer risk assessment,” *J Digit Imaging*, vol. 21, pp. 145–152, 2008.

[17] A. Manduca, M. Carston, J. Heine, C. Scott, V. Pankratz, K. Brandt, T. Sellers, C. Vachon, and J. Cerhan, “Texture features from mammographic images and risk of breast cancer,” *Cancer Epidemiol Biomarkers Prev*, vol. 18, pp. 837–845, 2009. [Online]. Available: http://dx.doi.org/10.1158/1055-9965.EPI-08-0631

[18] L. Häberle, F. Wagner, P. A. Fasching, S. M. Jud, K. Heusinger, C. R. Loehberg, A. Hein, C. M. Bayer, C. C. Hack, M. P. Lux *et al.*, “Characterizing mammographic images by using generic texture features,” *Breast Cancer Res*, vol. 14, no. 2, p. R59, 2012.

[19] M. Nielsen, G. Karemore, M. Loog, J. Raundahl, N. Karssemeijer, J. D. M. Otten, M. A. Karsdal, C. M. Vachon, and C. Christiansen, “A novel and automatic mammographic texture resemblance marker is an independent risk factor for breast cancer,” *Cancer Epidemiol*, vol. 35, pp. 381–387, 2011.

[20] J. J. Heine, C. G. Scott, T. A. Sellers, K. R. Brandt, D. J. Serie, F.-F. Wu, M. J. Morton, B. A. Schueler, F. J. Couch, J. E. Olson, V. S. Pankratz, and C. M. Vachon, “A novel automated mammographic density measure and breast cancer risk,” *J Natl Cancer Inst*, vol. 104, pp. 1028–1037, 2012. [Online]. Available: http://dx.doi.org/10.1093/jnci/djs254

[21] M. Nielsen, C. M. Vachon, C. G. Scott, K. Chernoff, G. Karemore, N. Karssemeijer, M. Lillholm, and M. A. Karsdal, “Mammographic texture resemblance generalizes as an independent risk factor for breast cancer,” *Breast Cancer Res*, vol. 16, p. R37, 2014. [Online]. Available: http://dx.doi.org/10.1186/bcr3641

[22] Y. Zheng, B. M. Keller, S. Ray, Y. Wang, E. F. Conant, J. C. Gee, and D. Kontos, “Parenchymal texture analysis in digital mammography: A

fully automated pipeline for breast cancer risk assessment," *Med Phys*, vol. 42, no. 7, pp. 4149–4160, 2015.

[23] M. Ranzato, C. S. Poultney, S. Chopra, and Y. LeCun, "Efficient learning of sparse representations with an energy-based model," in *Advances in Neural Information Processing Systems*, 2007, pp. 1137–1144.

[24] M. Ranzato, F. J. Huang, Y.-L. Boureau, and Y. LeCun, "Unsupervised learning of invariant feature hierarchies with applications to object recognition," in *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*. IEEE, 2007, pp. 1–8.

[25] K. Petersen, K. Chernoff, M. Nielsen, and A. Ng, "Breast density scoring with multiscale denoising autoencoders," in *Proc. Sparsity Techniques in Medical Imaging 2012, in conjunction with MICCAI 2012*, 2012.

[26] K. Petersen, M. Nielsen, P. Diao, N. Karssemeijer, and M. Lillholm, "Breast tissue segmentation and mammographic risk scoring using deep learning," in *Breast Imaging: 12th International Workshop, IWDM 2014*, ser. Lecture Notes in Computer Science, H. Fujita, T. Hara, and C. Muramatsu, Eds. Springer, 2014, vol. 8539, pp. 88–94.

[27] W. He, A. Juette, E. R. Denton, A. Oliver, R. Martı, and R. Zwiggelaar, "A review on automatic mammographic density and parenchymal segmentation," *International Journal of Breast Cancer*, 2015.

[28] R. Sivaramakrishna, N. A. Obuchowski, W. A. Chilcote, and K. A. Powell, "Automatic segmentation of mammographic density," *Academic Radiology*, vol. 8, no. 3, pp. 250–256, 2001.

[29] A. Torrent, A. Bardera, A. Oliver, J. Freixenet, I. Boada, M. Feixes, R. Marti, X. Llado, J. Pont, E. Perez, S. Pedraza, and J. Marti, "Breast Density Segmentation: A Comparison of Clustering and Region Based Techniques," in *IWDM '08: Proceedings of the 9th international workshop on Digital Mammography*. Berlin, Heidelberg: Springer-Verlag, 2008, pp. 9–16.

[30] R. J. Ferrari, R. M. Rangayyan, R. A. Borges, and A. F. Frère, "Segmentation of the fibro-glandular disc in mammograms using gaussian mixture modelling," *Med Biol Eng Comput*, vol. 42, pp. 378–387, 2004.

[31] R. Highnam and M. Brady, *Mammographic Image Analysis*. Kluwer Academic Publishers, 1999.

[32] S. van Engeland, P. R. Snoeren, H. Huisman, C. Boetes, and N. Karssemeijer, "Volumetric breast density estimation from full-field digital mammograms," *IEEE Trans Med Imaging*, vol. 25, pp. 273–282, 2006.

[33] J. Byng, N. Boyd, E. Fishell, R. Jong, and M. Yaffe, "Automated analysis of mammographic densities," *Physics in Medicine and Biology*, vol. 41, no. 5, p. 909, 1996.

[34] Z. Huo, M. L. Giger, D. E. Wolverton, W. Zhong, S. Cumming, and O. I. Olopade, "Computerized analysis of mammographic parenchymal patterns for breast cancer risk assessment: feature selection," *Med Phys*, vol. 27, pp. 4–12, 2000.

[35] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 8, pp. 1798–1828, 2013.

[36] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Networks*, vol. 61, pp. 85–117, 2015.

[37] Y. Bengio, "Learning deep architectures for AI," *Foundations and trends® in Machine Learning*, vol. 2, no. 1, pp. 1–127, 2009.

[38] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, pp. 2278–2324, 1998. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=726791

[39] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, "Learning hierarchical features for scene labeling," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, pp. 1915–1929, 2013.

[40] F. Ning, D. Delhomme, Y. LeCun, F. Piano, L. Bottou, and P. E. Barbano, "Toward automatic phenotyping of developing embryos from videos," *Image Processing, IEEE Transactions on*, vol. 14, no. 9, pp. 1360–1371, 2005.

[41] D. Ciresan, A. Giusti, L. M. Gambardella, and J. Schmidhuber, "Deep neural networks segment neuronal membranes in electron microscopy images," in *Advances in Neural Information Processing Systems*, 2012, pp. 2843–2851.

[42] S. C. Turaga, J. F. Murray, V. Jain, F. Roth, M. Helmstaedter, K. Briggman, W. Denk, and H. S. Seung, "Convolutional networks can learn to generate affinity graphs for image segmentation," *Neural Computation*, vol. 22, no. 2, pp. 511–538, 2010.

[43] D. Wei, B. Sahiner, H. Chan, and N. Petrick, "Detection of masses on mammograms using a convolutional neural network," in *Acoustic, Speech and Signal Processing*, vol. 5, 1995, pp. 3483–3486.

[44] P. Fonseca, J. Mendoza, J. Wainer, J. Ferrer, J. Pinto, J. Guerrero, and B. Castaneda, "Automatic breast density classification using a convolutional neural network architecture search procedure," in *SPIE Medical Imaging*. International Society for Optics and Photonics, 2015, pp. 941 428–941 428.

[45] A. R. Jamieson, R. Alam, and M. L. Giger, "Exploring deep parametric embeddings for breast cadx," in *SPIE Medical Imaging*. International Society for Optics and Photonics, 2011, pp. 79 630Y–79 630Y.

[46] A. R. Jamieson, K. Drukker, and M. L. Giger, "Breast image feature learning with adaptive deconvolutional networks," in *SPIE Medical Imaging*. International Society for Optics and Photonics, 2012, pp. 831 506–831 506.

[47] V. Jain, J. F. Murray, F. Roth, S. Turaga, V. Zhigulin, K. L. Briggman, M. N. Helmstaedter, W. Denk, and H. S. Seung, "Supervised learning of image restoration with convolutional networks," in *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*. IEEE, 2007, pp. 1–8.

[48] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, pp. 504–507, 2006.

[49] G. E. Hinton, "Learning multiple layers of representation," *Trends in cognitive sciences*, vol. 11, no. 10, pp. 428–434, 2007.

[50] P. Vincent, H. Larochelle, Y. Bengio, and P. A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *International Conference on Machine Learning*, 2008, pp. 1096–1103.

[51] J. Masci, U. Meier, D. Cireşan, and J. Schmidhuber, "Stacked convolutional auto-encoders for hierarchical feature extraction," in *Artificial Neural Networks and Machine Learning–ICANN 2011*. Springer, 2011, pp. 52–59.

[52] Q. Le, M. Ranzato, R. Monga, M. Devin, K. Chen, G. Corrado, J. Dean, and A. Ng, "Building high-level features using large scale unsupervised learning," in *International Conference in Machine Learning*, 2012.

[53] T. Lindeberg, *Scale-Space Theory in Computer Vision*. Dordrecht, the Netherlands: Kluwer Academic Publishers, 1994.

[54] L. Florack, "A spatio-frequency trade-off scale for scale-space filtering," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 22, no. 9, pp. 1050–1055, 2000.

[55] B. A. Olshausen *et al.*, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, no. 6583, pp. 607–609, 1996.

[56] H. Lee, C. Ekanadham, and A. Y. Ng, "Sparse deep belief net model for visual area V2," in *Advances in Neural Information Processing Systems*, 2008, pp. 873–880.

[57] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *International Conference on Machine Learning*, 2010, pp. 807–814.

[58] J. E. Olson, T. A. Sellers, C. G. Scott, B. A. Schueler, K. R. Brandt, D. J. Serie, M. R. Jensen, F.-F. Wu, M. J. Morton, J. J. Heine *et al.*, "The influence of mammogram acquisition on the mammographic density and breast cancer association in the mayo mammography health study cohort," *Breast Cancer Res*, vol. 14, no. 6, p. R147, 2012.

[59] G. Montavon, G. Orr, and M. K, Eds., *Neural Networks: Tricks of the Trade*. Springer, 2012, vol. 7700.

[60] L. M. J. Florack, B. M. ter Haar Romeny, M. A. Viergever, and J. J. Koenderink, "The Gaussian scale-space paradigm and the multiscale local jet," *Int J Comput Vis*, vol. 18, pp. 61–75, 1996.

[61] C. Nickson, Y. Arzhaeva, Z. Aitken, T. Elgindy, M. Buckley, M. Li, D. R. English, and A. M. Kavanagh, "Autodensity: an automated method to measure mammographic breast density that predicts breast cancer risk and screening outcomes." *Breast Cancer Res*, vol. 15, no. 5, p. R80, 2013. [Online]. Available: http://dx.doi.org/10.1186/bcr3474

[62] B. M. Keller, J. Chen, D. Daye, E. F. Conant, and D. Kontos, "Preliminary evaluation of the publicly available laboratory for breast radiodensity assessment (libra) software tool: comparison of fully automated area and volumetric density measures in a case–control study with digital mammography," *Breast Cancer Research*, vol. 17, no. 1, pp. 1–17, 2015.

[63] R. Wagner, M. Thom, R. Schweiger, G. Palm, and A. Rothermel, "Learning convolutional neural networks from few samples," in *Neural Networks (IJCNN), The 2013 International Joint Conference on*. IEEE, 2013, pp. 1–7.

[64] S. Tan and B. Li, "Stacked convolutional auto-encoders for steganalysis of digital images," in *Asia-Pacific Signal and Information Processing Association, 2014 Annual Summit and Conference (APSIPA)*. IEEE, 2014, pp. 1–4.

[65] T. L. Paine, P. Khorrami, W. Han, and T. S. Huang, "An analysis of unsupervised pre-training in light of recent advances," *arXiv preprint arXiv:1412.6597*, 2014.

[66] A. Droniou and O. Sigaud, "Gated autoencoders with tied input weights," in *International Conference on Machine Learning*, 2013, pp. 154–162.

[67] A. Coates, "Demystifying unsupervised feature learning," Ph.D. dissertation, Stanford University, 2012.