Discriminating solitary cysts from soft tissue lesions in mammography using a pretrained deep convolutional neural network

Thijs Kooi,[a] Bram van Ginneken, and Nico Karssemeijer
Department of Radiology and Nuclear Medicine, RadboudUMC, Geert Grooteplein Zuid 10, Nijmegen 6535, The Netherlands

Ard den Heeten
Department of Radiology, Academic Medical Center Amsterdam, P.O. Box 22660, DD Amsterdam 1100, The Netherlands

(Received 23 June 2016; revised 16 December 2016; accepted for publication 7 January 2017; published 16 March 2017)

Purpose: It is estimated that 7% of women in the western world will develop palpable breast cysts in their lifetime. Even though cysts have been correlated with risk of developing breast cancer, many of them are benign and do not require follow-up. We develop a method to discriminate benign solitary cysts from malignant masses in digital mammography. We think a system like this can have merit in the clinic as a decision aid or complementary to specialized modalities.
Methods: We employ a deep convolutional neural network (CNN) to classify cyst and mass patches. Deep CNNs have been shown to be powerful classifiers, but need a large amount of training data for which medical problems are often difficult to come by. The key contribution of this paper is that we show good performance can be obtained on a small dataset by pretraining the network on a large dataset of a related task. We subsequently investigate the following: (a) when a mammographic exam is performed, two different views of the same breast are recorded. We investigate the merit of combining the output of the classifier from these two views. (b) We evaluate the importance of the resolution of the patches fed to the network. (c) A method dubbed tissue augmentation is subsequently employed, where we extract normal tissue from normal patches and superimpose this onto the actual samples aiming for a classifier invariant to occluding tissue. (d) We combine the representation extracted using the deep CNN with our previously developed features.
Results: We show that using the proposed deep learning method, an area under the ROC curve (AUC) value of 0.80 can be obtained on a set of benign solitary cysts and malignant mass findings recalled in screening. We find that it works significantly better than our previously developed approach by comparing the AUC of the ROC using bootstrapping. By combining views, the results can be further improved, though this difference was not found to be significant. We find no significant difference between using a resolution of 100 versus 200 micron. The proposed tissue augmentations give a small improvement in performance, but this improvement was also not found to be significant. The final system obtained an AUC of 0.80 with 95% confidence interval [0.78, 0.83], calculated using bootstrapping. The system works best for lesions larger than 27 mm where it obtains an AUC value of 0.87.
Conclusion: We have presented a computer-aided diagnosis (CADx) method to discriminate cysts from solid lesion in mammography using features from a deep CNN trained on a large set of mass candidates, obtaining an AUC of 0.80 on a set of diagnostic exams recalled from screening. We believe the system shows great potential and comes close to the performance of recently developed spectral mammography. We think the system can be further improved when more data and computational power becomes available. [https://doi.org/10.1002/mp.12110]

Key words: breast cancer, computer aided diagnosis, deep learning, solitary cysts, transfer learning

1. INTRODUCTION

It is estimated that 7% of women in the western world will develop palpable breast cysts in their lifetime. Even though cysts have been correlated with risk of developing breast cancer,[1] many of them are benign and do not require follow-up. On mammography, benign cysts and solid lesion can be difficult to discriminate and consequently, many women are being recalled unnecessarily for a second diagnostic exam or core needle biopsy. Literature suggest that 20%[2] to 37%[3] of recalls can be attributed to benign solitary cysts. False positives have been shown to cause severe psychological stress up to 3 years after diagnosis, sometimes as high as women diagnosed with breast cancer.[4]

To better differentiate between findings, ultrasound is commonly used in diagnostic examinations, but this is typically not available during screening. In a recent study, Erhard et al.[2] employed spectral mammography, an extension of

mammography utilizing two energy thresholds with a single exposure, which allows measuring attenuation and consequently cystic and lesion volume accurately. Two features are computed, the first based on the ratio of cystic against lesion volume and a second one measuring the cystic diameter. A linear discriminant analysis (LDA) is subsequently used for classification. Their method obtained an area under the ROC curve (AUC) of 0.88 with a median specificity of 56% at 99% sensitivity. The authors estimate the employment of their method in the clinic would result in a reduction of 32% of recalls based on well-defined solitary lesions and an overall reduction in recall of 6%.[2]

Aside from specialized modalities, computer-aided detection (CADe) or diagnosis (CADx)[5,6] systems are being developed to aid readers of medical images. Mammography has so far been on the forefront of this development, where CAD has FDA approval in the US, is reimbursed and is already widely applied as a second reader.[7,8] In spite of this, there is still much debate regarding the efficacy of current technology, with some studies suggesting CADe for masses increases neither specificity nor sensitivity.[9,10] Additionally, research suggests most human error does not stem from oversight but rather misinterpretation.[8] Methods that focus on classification rather than detection could therefore have far more merit in the clinic.

CAD systems generally rely on machine learning methods, which have seen a surge of new interest with the advent of deep learning[11,12] resulting in several landmark studies.[11,13,14] Techniques are slowly but steadily making their way into medical image analysis.[15–17] The key strength of deep learning lies in the fact that feature hierarchies are learned from data rather than defined manually, preventing any bias imposed by the engineer. The consensus seems to be that the biggest downside is the supervised nature of the training procedure,[18] warranting large amounts of annotated data for each task, which for medical problems may prove to be a big hurdle. Learning features on a (mildly) related problem has been shown to improve performance in various medical tasks.[19–23]

In this paper, we present a CADx system to discriminate benign cysts from solid lesion in regular digital mammography. The key contribution is showing that a deep Convolutional Neural Network (CNN) can be used effectively on a small dataset by pretraining the network on other data. Contrary to related work,[23,24] the network is not pretrained on natural images, but on a large dataset of soft tissue lesion candidates from a set of screening mammograms. Features are extracted from the network and a second classifier is trained for the diagnosis task. The method is compared to the only previously developed method for this problem, where a model of the cysts and solid lesions was applied to derive two descriptors invariant to surrounding tissue.

During a mammographic examination, two views of each breast are typically recorded and when radiologists evaluate a case, the information in both is taken into account. We show that by combining these two views, better classification performance can be obtained. We subsequently show that by augmenting samples with normal tissue from normal areas in the breast higher performance can be obtained. Lastly, we combine the features extracted from the CNN with the previously developed method to obtain our final results.

The method is evaluated on a large dataset of roughly 1000 malignant masses and 600 cysts. Typical examples of masses and cysts in our dataset are shown in Fig. 1. To the best of our knowledge, this is the best performance reported in literature for this problem in digital mammography and the first method that investigates transfer between related medical tasks in deep CNNs.

The rest of this paper is divided into five sections. In the section 2, we will provide an overview of the data. Section 3 A will give details on the CNN setup and how data augmentation is applied, followed by a brief description of the previously published method which we compare the CNN against. In section 4, we provide experimental details and results, followed by a discussion in section 5 and conclusion in section 6.

2. DATA

We make use of two different datasets. The first dataset concerns a large collection of screening mammograms obtained from a screening program in The Netherlands (*bevolking-sonderzoek midden-west*) and recorded using a Hologic Selenia digital mammography system at a resolution of 70 micron. All tumors are biopsy-proven malignancies and annotated under the supervision of a certified radiologist. The data were randomly split into a training and validation set on a patient level. This dataset is used to pretrain the CNN.

The second set was used for retraining and producing the final results, which were obtained using nested cross-validation with eight inner and outer folds with all data split on a patient level. It consists of diagnostic exams of women who were recalled in screening for a suspicious mass lesion. Again, all malignant lesions are biopsy proven. Images were recorded using a GE Senograph 2000D(S) at an original resolution of 100 micron. All soft tissue lesions used in this study were either marked as ductal carcinoma in situ, invasive ductal carcinoma, or invasive lobular carcinoma. Masses marked as lobular carcinoma in situ were removed from the dataset. All cysts in the dataset were classified as solitary cysts by experienced radiologists. All images with multiple cysts were removed from the dataset, even though they were recalled in screening. The dataset contained three cases with a breast implant, which were removed from the set as well (Table I).

In both datasets, we extracted patches of 260 $\times$ 260 at 200 micron, or 5.2 cm per patch (unless mentioned otherwise in the experiments section). The center of each patch was taken as the mean x and y values of a contour drawn by research assistants under the supervision of experienced radiologists.

3. METHODS

3.A. Deep convolutional neural network

Many deep learning applications enjoy gains in terms of improved accuracy or lower training times by employing a

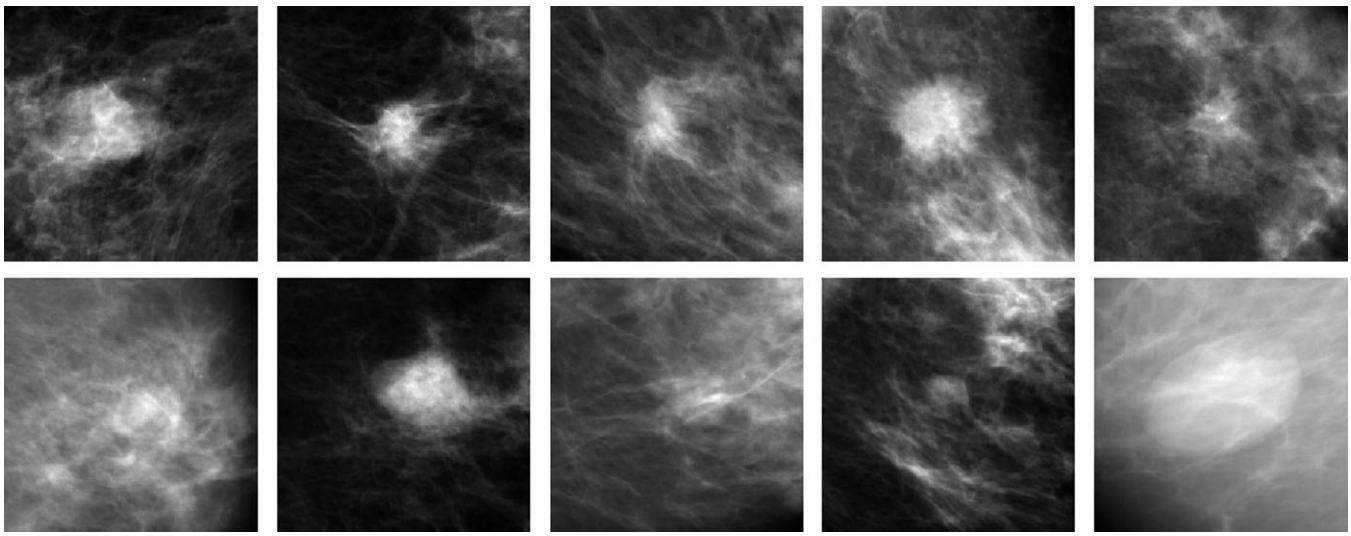

FIG. 1. Typical examples of malignant masses (top row) and benign solitary cysts (bottom row) in our dataset. Each patch captures 5.2 cm or 260 pixels at a resolution of 200 micron. Cysts are fluid-filled sacs and part of normal tissue and malignant masses are signs of cancer and require follow-up examination. Unfortunately, they are often difficult to distinguish during mammography screening meaning cancers can be missed or women are recalled unnecessarily for a follow-up examination. We are developing a computer-aided diagnosis (CADx) method to discriminate malignant masses from benign cysts in mammography.

network trained on large annotated datasets such as ImageNet[25] and subsequently applying it off-the shelf or finetuning it for the particular task,[26,27] in particular when training data for the actual task is scarce. A similar strategy was also used for several medical problems.[19–23]

An arguable downside of using a network trained on natural images is that the input data are significantly different. Natural images often have three channels, whereas medical images can have any format, ranging from gray scale to volumetric, meaning channels need to be copied thereby wasting network parameters. In addition, learned invariance properties may not transfer well or may not have been learned from the source dataset. For instance, object classes in natural images are typically invariant to linear intensity changes, whereas many medical imaging problems are not. Conversely, natural objects are generally not rotation invariant and tumors in mammography are to some extent. Learning features on a large dataset of a more related task could therefore make transfer easier.

Although annotated data of abnormalities in many medical imaging problems is often difficult to come by, normal data can be far more ubiquitous. A possible approach to initialize a network would be to simply train a set of stacked Autoencoders on randomly selected normal samples and subsequently finetune the network for the real classification task. The actual information in normal samples, however, may be low and many weights in the network will be dedicated to represent structure without any discriminative power for the actual problem. Instead, we proposed to train fully supervised on a large set of mass candidates: patches found by a candidate detector to resemble masses, extract feature representations from the network, and learn a new classifier for diagnosis.

To this end, we used a large dataset of unprocessed screening mammograms. All images were log-transformed and inverted and the breast and pectoral muscle area were segmented. We subsequently extracted five texture features, two looking for the center of a focal lesion, two designed to spot spiculae, characteristic of malignant soft tissue lesions, and a last feature capturing the optimal response in scale-space.[28] An ensemble of five multilayered perceptrons (MLPs) was then trained on this feature set. Normal pixels for the training set were sampled randomly from normal parts in the segmented breast area, and positive datapoints were sampled densely from a circle of constant size, inside each annotated malignant mass. Pixels in all images were classified using the learned ensemble to form a likelihood image on which we performed nonmaximum suppression to generate a set of candidate locations for each image. Centered at each location we extracted patches, which are then used to train the deep CNN. Figure 2 shows examples of true and false positives the network is trained with.

Rather than retraining the full network, we treated the CNN as a feature extractor and simply extracted latent representations from the network by feeding cyst and mass patches and trained a shallow nonlinear classifier on this feature set. This has a main advantage that the experiments can be run in cross-validation in reasonable time and that the hyperparameters of contemporary ('shallow') classifiers such as gradient boosted trees (GBT) are easier and faster to tune. The latent representation can be accessed from any part in the network, but the dimensionality increases closer to the input patch, imposing computational and memory problems. Instead, we focused on the last three layers. Figure 3 shows an illustration of the network and the layers at which we extracted features.

An arguable shortcoming of the deep neural network models is that prior knowledge, such as knowledge about invariances to nuisance factors in the classification problem, other than translation, is difficult to encode into the architecture. Rather, CNNs are typically fed data that are transformed in such a way that it reflects as many possible variations that we

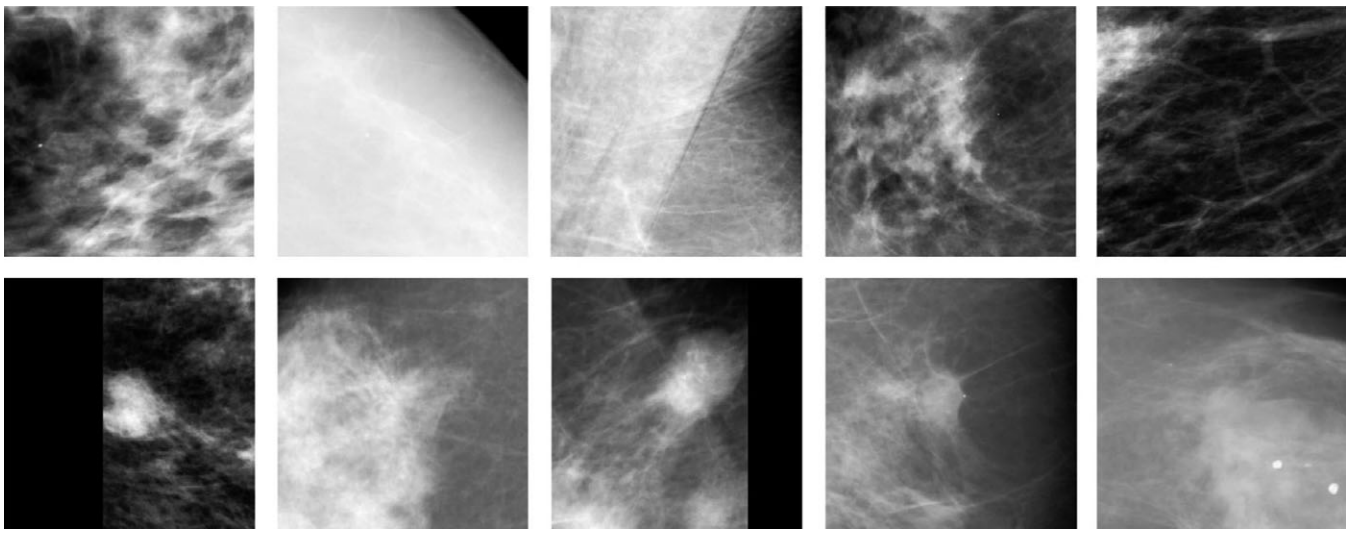

FIG. 2. Examples of false-positive (top row) and true-positive (bottom row) lesion candidates. As the set of solitary cysts and malignant masses is relatively small compared to the amount of data deep neural networks typically need, we first train a network on a big set of mass candidates taken from a large database of screening mammograms.

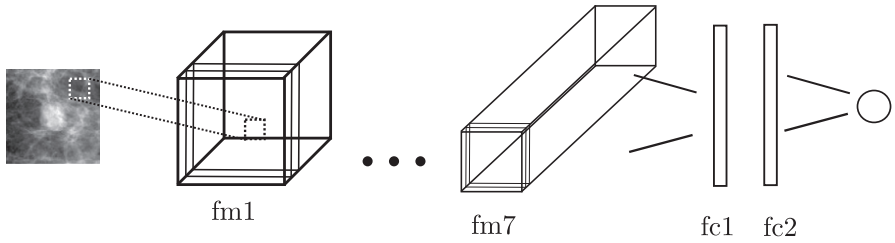

FIG. 3. Illustration of the deep Convolutional Neural Network (CNN) employed. The network is trained on a large dataset of potential mass regions from screening mammograms (see Fig. 2). Solitary cyst and malignant mass patches from a diagnostic dataset are subsequently presented to the network and features are extracted from the hidden layers and used for the diagnosis task. The network employed is similar to the one in Ref.[33,34]. For brevity, only the first feature maps and fully connected layers used are shown. The exact architecture is described in section 4. Abbreviations fc1, fc2, and fm7 refer to the fully connected layers, feature maps, and their index.

expect to see in the image in the hope that the network learns an invariant representation with respect to these nuisance factors. The process of adding training samples this way is known as *data augmentation*. For natural images, simple scaling, translation, and color transformations are typically applied, but for medical data, different sources of variation are present. One of such sources is variation in the amount of tissue surrounding a tumor. All other things being equal, the same tumor will look different on the mammogram if fibroglandular is added or removed. To simulate different forms of occluding tissue, we proposed to use *tissue augmentation*, where parenchymal patterns from different images are added to the patch in a physically plausible fashion.

3.B. Tissue augmentations

To perform tissue augmentation, we manually selected 200 patches from normal regions in mammograms of normal breasts recorded with the same detector. In doing so, we made sure the patch was sufficiently far from the breast boundary and that the full patch contained tissue (i.e., no pectoral muscle or boundaries of the breast). To perform realistic augmentations, we make use of the following model of pixel value $I(\mathrm{x})$ at 2D location x:

$$I(\mathrm{x}) = e^{-h(\mathrm{x})\mu(\mathrm{x})} \tag{1}$$

with h the height of the tissue and μ the linear attenuation coefficient. By taking the log and adding the augmentation patch also in log space, super scripted a, to source image s we get

$$\log[I(\mathrm{x})] = -h^s(\mathrm{x})\mu^s(\mathrm{x}) + -h^a(\mathrm{x})\mu^a(\mathrm{x})$$

Therefore, physically plausible augmentations entail a simple addition. To prevent unrealistically thick patches, we introduce a blending factor α that governs the amount of tissue added to the source patch:

$$\log[I_t] = \log[I_s] + \alpha\log[I_a] \tag{2}$$

Figure 4 shows several examples of normal patches (top row) which are added to the first image in the bottom row to generate augmented patches shown in the rest of the bottom row. Each patch in the training set was blended with eight randomly selected patches from the pool of normals. We have chosen to apply the tissue augmentations in the fine tuning

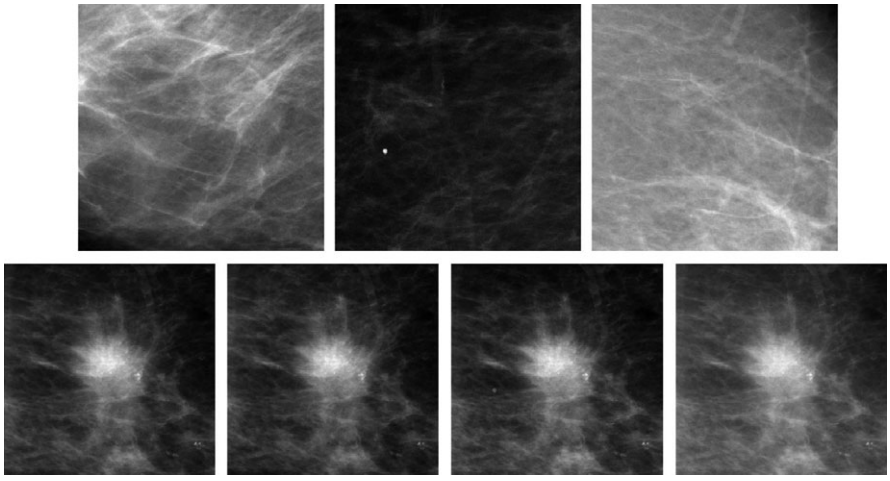

FIG. 4. Deep CNNs are largely problem agnostic methods. To make them learn the right invariance properties, data augmentation is typically performed, where patches are transformed in such a way that they represent all possible variation in the data. We propose a data augmentation method dubbed tissue augmentation, where we randomly select normal tissue from normal areas in the breast and superimpose these over mass and cyst patches to simulate different amounts of parenchymal tissue surrounding the lesions. (Top row) Normal patches that are superimposed on the leftmost patch in the bottom row generating patch 2–4 in the bottom row. We used a blending factor $\alpha = 0.4$

stage and not to the pretrained network, such that the effect of the α parameter can be evaluated more easily.

3.C. Contrast features

To the best of our knowledge, no methods have been published that tackle this specific problem in mammography. We therefore compare the system to our own previous work, as described in Ref.[29] which we briefly repeat here. The method is based on contrast features, popular descriptors of candidate findings in mammography prior to the deep learning paradigm, which were designed to be invariant to tissue surrounding a potential tumor. This type of feature typically relies on a segmentation of the candidate, for which we use the segmentation algorithm proposed by Timp et al.,[30] which was shown to be superior to other methods on their particular feature set and dataset.

To derive the two contrast features, a simple model of the lesion f was proposed:

$$\log[I(x,y)] = \begin{cases} F_z(x,y) + \epsilon & \text{if } (x,\ y) \text{ lies inside the 2D projection of } f \\ \epsilon & \text{else} \end{cases} \tag{3}$$

where $I(x,\ y)$ indicates the image value of the unprocessed mammogram at location $(x,\ y)$, $\epsilon \sim P(\epsilon)$ is the integral along the z-axis of the nuisance term, the surrounding tissue, we are trying to ignore, coming from some undefined distribution. $F_z\ (x,\ y) = \int f(x,\ y,\ z)dz$ denotes the z-integral of the lesion f we are trying to describe. The z-axis is chosen to be parallel to the direction of x-ray quanta. This model assumes the 2D segmentation of the projected lesion is correct and that the tumor grows in such a way that the distribution of tissue in the surrounding region is the same as above and below it.

Two descriptors were subsequently derived to be invariant to surrounding tissue and scaling. Assuming cysts are easily compressed during the recording of a mammogram and therefore behave like ellipsoids, and masses are hard and behave like spheres, gives the following two features:

$$\hat{\mathbb{E}}[F_z] = \frac{\mathbb{E}[F_z + \epsilon] - \mathbb{E}[\epsilon]}{r} \tag{4}$$

and

$$\hat{Var}[F_z] = \frac{Var[F_z + \epsilon] + Var[\epsilon] - 2Cov[F_z + \epsilon, \epsilon]}{r^2} \tag{5}$$

4. EXPERIMENTS

4.A. Deep CNN learning settings

The CNN was implemented in Theano.[31] We used a VGG-like network architecture[32] that was similar to the one employed in Kooi et al.[33,34] with two additional convolutional layers. The model employs 3×3 kernels in seven convolutional layers with {16, 16, 32, 32, 64, 128, 128} feature maps, pooling on {2, 1, 2, 1, 2, 1, 2} and ReLU units in all layers, except for the classification layer. Two fully connected layers with 300 units each were added. We have experimented with deeper networks and more feature maps but did not see an improvement in performance on the validation set. We employed a binary cross-entropy loss which was minimized using ADAM[35] and a learning rate of 0.00005. Dropout[36] was applied to all fully connected layers with $P = 0.5$. We added an L2 norm with a weight of 0.00005 to all layers. Weights were initialized using the MSRA weight filler.[37] Early stopping was used using the AUC on the validation set.

As the dataset is heavily imbalanced, we generated two separate sets: one with all normal samples and one with all malignant masses. We iterate through the set of normals, reading chunk by chunk from disk and hold the set of positives in host RAM. For each minibatch of normals, we randomly sample an equal amount malignant masses from the

set in host RAM to maintain a 50/50 class ratio. The model was trained for 5 days on a Titan X, 12 GB GPU. For our pretrained network, we augmented each patch 16 times with scaling, translation and all eight reflective symmetry permutations.

Patches were scaled between 0 and 1, when employing tissue augmentations the scaling factors were multiplied by $1 + \alpha$. As some candidates occur at the border of the breast, we padded each mammogram with zeros.

4.B. Top layer learning settings

For both the manual features and the features extracted from the CNN, we employed a gradient boosting tree (GBT) classifier[38] with a binomial deviance loss function:

$$\mathcal{L}(y, h(\mathrm{x})) = \log[1 + \exp(-2yh(\mathrm{x}))] \quad (6)$$

a tighter upper bound on the zero-one loss and less susceptible to outliers than the exponential or squared loss. We employed nested cross-validation with eight inner folds on the maximum depth and shrinkage and 16 outer folds for testing. We used the square root heuristic for the maximum number of features. Data were split on a patient level to prevent any bias in both inner and outer fold. In all settings, 100 estimators were used. A weight was used inversely proportional to the class ratio to account for imbalance.

During screening, two images are typically recorded of each breast: a top-down view (CC) and sideways view (MLO), which gives the radiologists additional information. To harness this information, we also experimented with combining the classifiers posterior of the two different views. We evaluated two simple rules: mean and max. If only one view is present or the lesion is only annotated in one of the two views, only this view was used.

4.C. Results

As was done in our previous work,[29] we first compared the tissue normalized descriptors to two similar naive contrast features: A simple estimate of the mean

$$\mathbb{E}[F_z + \epsilon] - \mathbb{E}[\epsilon] \quad (7)$$

and variance:

$$Var[F_z + \epsilon] - Var[\epsilon] \quad (8)$$

Figure 5 shows the ROCs obtained with the different sets of contrast features. Figure 6 shows results obtained when combining views using the max rule. 95% Confidence intervals and P-values were computed using bootstrapping[39] with 2000 bootstraps in all settings. We found no significant difference ($P = 0.4456$, mean difference 0.0017 with 95% confidence interval $[-0.0306, 0.0325]$) between the unnormalized and normalized features when using only a single view, but did find a significant difference when combining them using the max rule ($P = 0.0286$).

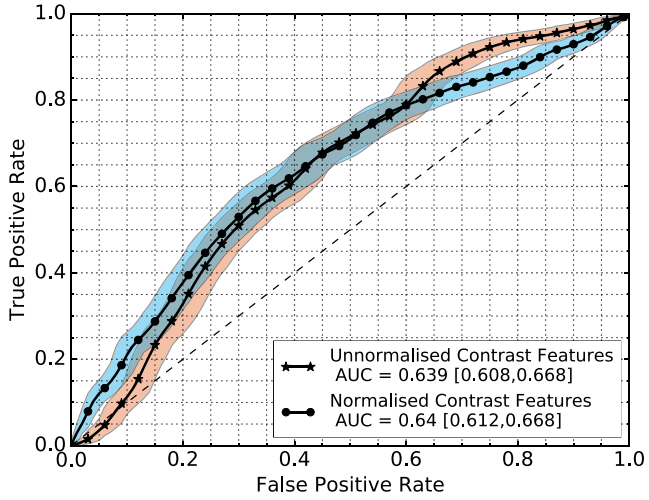

FIG. 5. ROC results plus a 95% confidence interval of the different sets of contrast features using only a single view. Normalized features refers to the features in Eqs. (4) and (5), and unnormalized features as defined in Eqs. (7) and (8) [Colour figure can be viewed at wileyonlinelibrary.com]

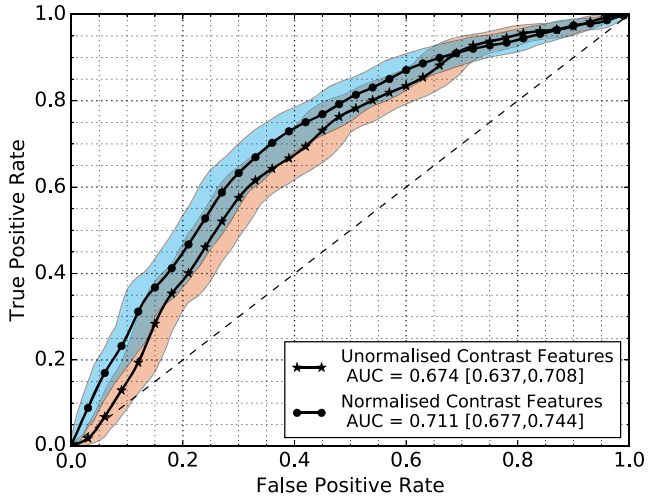

FIG. 6. ROC results plus a 95% confidence interval of the different sets of contrast features when combining CC and MLO views using the max rule. [Colour figure can be viewed at wileyonlinelibrary.com]

4.C.1. Effect of the depth

We subsequently investigated the optimal level at which features can be extracted. When extracting features deeper in the network, the size of the feature vector increases, thereby adding some computational burden, so if two depths perform equal in terms of classification performance, the smaller representation is still preferred. Additionally, it gives insight into how 'transferable' the latent representation is, i.e., if feature maps closer to the input give optimal performance, it would suggest the learned features are suitable for both tasks.

Table II shows the AUC values and a 95% confidence interval, obtained after retraining the network with features extracted from the final fully connected layer (fc2), the first fully connected layer (fc1), and the last feature maps (fm7).

TABLE I. Overview of dataset. We used two different sets: (a) A large dataset of screening mammograms containing normal exams and biopsy-proven malignant masses (i.e., no cysts or other benign abnormalities) that was used to learn the features from the CNN and (b) a set of diagnostic exams referred from screening containing biopsy-proven malignant masses and benign solitary cysts. For the first dataset, we used a separate validation set to optimize the CNN's hyperparameters, and the amount of training and validation samples are indicated as (# training/ #validation). As the second dataset is relatively small, we used nested cross-validation and hence no fixed validation or test set was employed.

Screening Exams (Hologic)	Images	Exams
Normal	(73102/21913)	(20000/8979)
Malignant masses	(1487/342)	(1000/205)
Diagnostic Exams (GE)		
Malignant masses	1108	586
Solitary Cysts	696	370
Total	1804	956

TABLE II. AUC values and 95% confidence interval obtained after extracting features from different layers in the network using patch resolution of 200 micron and retraining a classifier for the task at hand. Abbreviations fc and fm indicate the index of the fully connected layer and feature map, respectively. See Fig. 3 for an illustration. Image based refers to the results obtained after classifying every lesion individually and exam based refers to the results obtained after combining the output of the classifier for the CC and MLO of the exam.

Layer	Nr of features	Image based	Exam based
fc2	300	0.741 [0.717, 0.764]	0.748 [0.715, 0.780]
fc1	300	0.744 [0.720, 0.767]	0.752 [0.719, 0.783]
fm7	21632	0.767 [0.747, 0.79]	0.773 [0.742, 0.802]

Single and combined refers to using only a single view and combining two views, as described earlier. We have experimented with feature maps extracted deeper in the network but did not see a clear increase in performance. We found a significant difference between the performance of last feature maps (fm7) and the normalized features, both using single view and the combination ($P = 0.01$).

We did not find a clear difference between using the mean and max rule on this feature set and therefore only the max rule is shown in the table. Although there is an improvement when going deeper and combining views, we also did not find this to be significant ($P \gg 0.05$, mean difference between fm7 and fc2 on exam level: 0.0247 with 95% confidence interval [0.0476, −0.0025]) in any combination. Figure 7 shows the ROC curves comparing the single and combined classification with features extracted from the final feature maps.

4.C.2. Effect of resolution

In the next setting, we extracted the cyst and mass patches at 100 micron rather than 200, but kept using the same network trained at a resolution of 200. We used the fully convolutional approach described by Long et al.[40] to also extract the representation from the fully connected layers. This test can give insight into two things: (a) is there any possible additional information when going for a higher resolution and (b) what kind of features are actually learned by the network. Table III shows AUC values and 95% confidence intervals obtained after extracting features from different layers in the network, this time retrained on patches of 100 micron. Results show the higher resolution gives very marginal improvements, but they are not found to be statistically significant ($P \gg 0.05$ red mean difference between 0.002, 95% confidence interval [−0.0203, 0.0164]) and considering the far higher computational needs, we continued the rest of the experiments with the 200 micron patches.

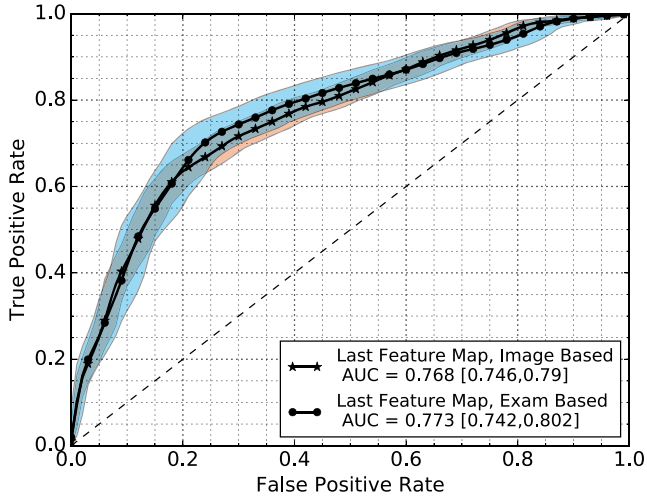

FIG. 7. ROC curves plus 95% confidence interval obtained after using the features extracted from the last feature map of the CNN. (final row in Table II) [Colour figure can be viewed at wileyonlinelibrary.com]

4.C.3. Effect of Mixing

We subsequently investigated the effect of the mixing parameter α. Table IV shows AUC values plus again a 95% confidence interval obtained after bootstrapping. Figure 8(a) shows the ROC curve obtained when applying the best α (last row in Table IV). Although we see an improvement using several values of α, we did not find these to be significant when comparing them to the performance without ($P = 0.9$ for the best value, 0.0470 95% confidence interval [0, 0.0240]).

4.C.4. Effect of combining with contrast features

In an attempt to further increase the performance of the system, we combined the normalized contrast features with the features extracted from the pretrained CNN by simply concatenating the feature vectors. For this, we used the value of α found to be optimal in the previous experiment. We obtained an AUC of 0.784 with 95% confidence interval [0.763, 0.805] for the single view setting and an AUC of 0.804 with 95% confidence interval [0.776, 0.832] when

TABLE III. AUC values and 95% confidence interval obtained after extracting features from different layers in the network using patch resolution of 100 micron and retraining a classifier for the task at hand. Abbreviations fc and fm indicate the index of the fully connected layer and feature map, respectively. See Fig. 3 for an illustration.

Layer	Nr of features	Image based	Exam based
fc2	86700	0.762 [0.74, 0.784]	0.775 [0.74, 0.805]
fc1	86700	0.765 [0.742, 0.787]	0.775 [0.744, 0.805]
fm7	107648	0.769 [0.746, 0.791]	0.777 [0.745, 0.806]

TABLE IV. Performance of the system after tissue augmentation, varying the blending factor α (see section 3 B). The second and third columns show the mean AUC plus a 95% confidence interval obtained after bootstrapping.

α	Image based	Exam based
0.8	0.762 [0.739, 0.785]	0.779 [0.748, 0.810]
0.72	0.764 [0.741, 0.786]	0.785 [0.753, 0.816]
0.63	0.761 [0.738, 0.784]	0.784 [0.752, 0.814]
0.55	0.762 [0.739, 0.785]	0.783 [0.750, 0.813]
0.47	0.757 [0.734, 0.78]	0.777 [0.745, 0.807]
0.38	0.736 [0.712, 0.76]	0.755 [0.722, 0.786]
0.3	0.73 [0.706, 0.753]	0.749 [0.716, 0.781]
0.22	0.733 [0.709, 0.757]	0.75 [0.716, 0.782]
0.13	0.757 [0.735, 0.780]	0.777 [0.745, 0.807]
0.05	0.775 [0.753, 0.797]	0.795 [0.764, 0.824]

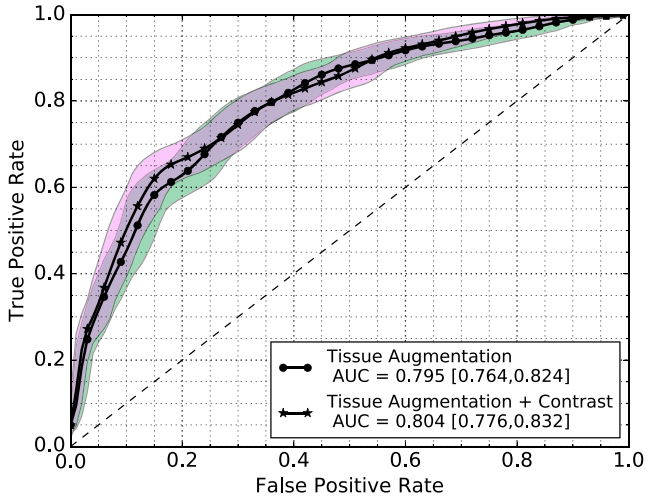

FIG. 8. ROC curves obtained after using the features from the last feature map and tissue augmentation with $\alpha = 0.05$ and the same configuration after combining it with the manually derived contrast features. We see small improvements by combining the methods but these were not found to be significant. [Colour figure can be viewed at wileyonlinelibrary.com]

combining views. Plots of the ROC curves are provided in Fig. 8.

4.C.5. Effect of size

Lastly, similar to Erhard et al.,[2] we split the performance of the system over different sizes of lesions. Results for the 200 micron case at fm7 are shown in Table V. The AUC for all lesions of size greater than 10 mm was 0.806 with 95% confidence interval [0.777, 0.834]. The images in Figure 9 show the cysts and masses that were found to be most difficult by the final system.

5. DISCUSSION

From the ROC curves in Fig. 6, we can see the normalizations as described in section 3 C to improve performance when combining views, but not for the single view case. This is contrary to what we found in previous work. Part of this can be explained by the fact that the dataset used in this study is more difficult and in previous work only images were used where two views were present.

It seems the improvement obtained when combining views is less when the method works better on a single view. We believe a big part of this can be explained by the fact that the closer you get to the optimal information a system gets out of a dataset, the more difficult it is to improve upon the results.

From the results in Table II, we can see marginal improvements in performance when extracting latent representations deeper in the network. We have tried extracting features closer to the input as well, but did not see an increase in performance. This suggests the tasks are related enough for easy feature transfer and is in contrast to results reported when transferring between natural images and medical images,[24] where retraining deeper has been shown to be beneficial for several tasks. A downside of our approach is that the dimensionality of the feature vector increases the deeper the features are extracted. We think retraining the full network could result in a minor performance increase, but due to the difficulty of running this in cross-validation, we have chosen not to do this.

Interestingly, the performance was roughly similar at 100 micron. As we did not see a very clear increase in performance either, we have chosen to use the downscaled version only. However, the roughly equal performance gives the idea that the features learned are not responding to any specific part but are more general texture descriptors useful for

TABLE V. AUC values with 95% confidence interval of the system for different lesion sizes. The second column indicates the amount of samples, with the amount of malignant masses and benign solitary cysts between brackets. The third column shows the performance for different lesion sizes, using the $\alpha = 0.05$, the fourth column the same system but with added contrast features.

Diameter range	Nr of samples	AUC	AUC with contrast
0–10 mm	196 (43/25)	0.753 [0.624, 0.87]	0.776 [0.658, 0.878]
10–13.5 mm	192 (110/66)	0.826 [0.760, 0.884]	0.784 [0.713, 0.850]
13.5–17 mm	162 (132/87)	0.714 [0.643, 0.781]	0.835 [0.782, 0.885]
17–20 mm	109 (111/65)	0.805 [0.729, 0.875]	0.782 [0.707, 0.851]
20–27 mm	163 (136/89)	0.818 [0.757, 0.873]	0.810 [0.751, 0.866]
>27 mm	134 (54/38)	0.866 [0.788, 0.932]	0.826 [0.736, 0.904]

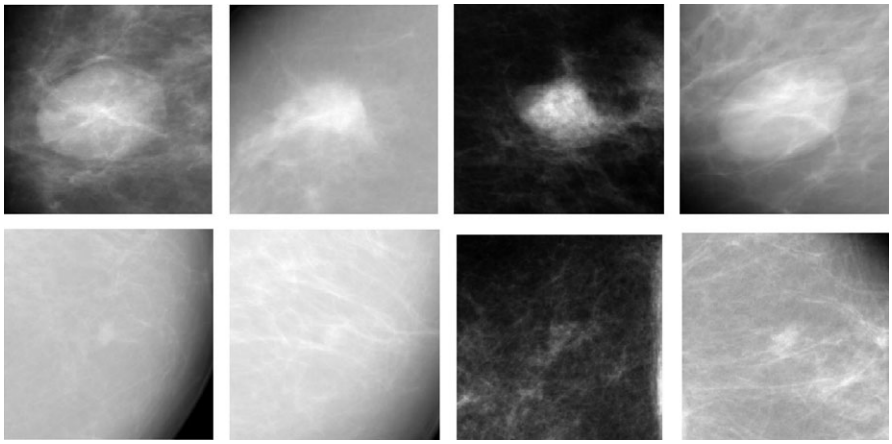

FIG. 9. (Top row) Cysts that were classified as most like masses by the final system in single view mode. (Bottom row) Malignant masses that were classified as most like solitary cysts by the final system in single view mode. Very large cysts and very small masses are most difficult for the system. We suspect these are simply underrepresented in the training set.

mammography, which we think is an interesting finding. In future work, we plan to explore training a network on mass candidates at 100 micron and subsequently retraining it for this task to see if the performance increases.

From the results in Table IV, we see the tissue augmentations give a small amount of improvement for large values of α and the most improvement over the baseline for an $\alpha = 0.05$, though none of these improvements were found to be significant. In this study, we have shown results for all values of the mixing coefficient to get some insight into its performance on a test set. When using the system, the parameter should be cross-validated, similar to the parameters of the classifiers. We do believe that this will still give an improvement in performance. Similarly, we found the scaling factors to be an important influence, for which we now choose one setting, but these can equally be cross-validated for better performance.

In recent years, several groups have started working on deriving networks invariant to basic geometric transformations[41–43] possibly obviating the need for geometric data augmentation. We feel the tissue augmentations proposed are particularly relevant, as it may prove to be difficult to derive networks invariant to this nuisance factor mathematically.

In comparison to a similar study performed with spectral mammography, a specialized modality, Erhard et al.[2] obtained an AUC of 0.88. Their set contained only 62 solid and 52 cystic lesions. From the solid lesions, 15 were benign. Similar to their study, we find the method works slightly better for larger lesions, though very small and very large lesions are most difficult for the system (see Fig. 9). We suspect these are simply underrepresented in the training set and as training data become more ubiquitous the performance will increase.

From the curves in Fig. 8, we can see that the performance already comes close to that reported by Erhard et al.[2] and that part of the lesions can be filtered out and do not need to be recalled. As with any study, the performance can be worse when tested on other data, as there is still some variation in the AUC indicated by the confidence intervals. However, we think the proposed deep CNN approach shows great potential and even better results could be obtained. For instance, by looking at Table III, we can see a very small increase in performance when using patches extracted at 100 micron. By using a network also pretraining on mass candidates at a resolution of 100 micron and subsequently using the features from this network, we suspect the results to be better and more information can be extracted from the patches. Unfortunately, the increased training time is currently still prohibitive to execute this. Lastly, the context of the mammogram has not been taken into account yet. When radiologists look at an exam, they will most likely not only consider a patch but also look at the image as a whole, which could result in better diagnosis.

6. CONCLUSION

In this paper, we have presented a computer-aided diagnosis (CADx) method to discriminate cysts from solid lesion in mammography using features from a deep CNN trained on a large set of mass candidates, obtaining an AUC of 0.8 on a set of mass candidates recalled from screening. We have compared the CNN-based system to our own previous work and only method published for this problem and have shown it outperforms this method. Contrary to related work investigating transfer,[23,24] we do not use pretrained networks on natural images, but use a large dataset from a more related task.

We have shown that by augmenting the patches with randomly sampled tissue from normal images, small improvements in performance can be obtained. The final system works best for lesions larger than 20 mm where it obtains an AUC of 0.866. The AUC of the system comes close to the AUC obtained with the recently proposed spectral mammography,[2] which is promising. We also believe that with more

data and more computational power, the performance can still be improved significantly.

ACKNOWLEDGMENTS

This research was funded by grant KUN 2012-5577 of the Dutch Cancer Society and supported by the Foundation of Population Screening Mid West.

CONFLICTS OF INTEREST

Bram van Ginneken is cofounder and shareholder of Thirona (Nijmegen, The Netherlands). Ard den Heeten is cofounder of Sigmascreening (Amsterdam, The Netherlands) and consultant of Volpara Health Technologies Ltd. (Wellington, New Zealand). Nico Karssemeijer is cofounder, shareholder, and director of ScreenPoint Medical BV (Nijmegen, The Netherlands), cofounder of Volpara Health Technologies Ltd. (Wellington, New Zealand), and QView Medical Inc. (Los Altos, CA).

[a]Author to whom correspondence should be addressed. Electronic mail: thijs.kooi@radboudumc.nl.

REFERENCES

1. Dixon JM, McDonald C, Elton R, Miller W. Risk of breast cancer in women with palpable breast cysts: A prospective study. *Lancet*. 1999;353:1742–1745.
2. Erhard K, Kilburn-Toppin F, Willsher P, et al. Characterization of cystic lesions by spectral mammography: Results of a clinical pilot study. *Invest Radiol*. 2016;51:340–347.
3. Sickles EA. Probably benign breast lesions: When should follow-up be recommended and what is the optimal follow-up protocol? *Radiology*. 1999;213:11–14.
4. Brodersen J, Siersma VD. Long-term psychosocial consequences of false- positive screening mammography. *Ann Fam Med*. 2013;112:106–115.
5. Kunio D. Computer-aided diagnosis in medical imaging: Historical review, current status and future potential. *Comput Med Imaging Graph*. 2007;31:198–211.
6. Van Ginneken B, Schaefer-Prokop C, Prokop M. Computer-aided diagnosis: How to move from the laboratory to the clinic. *Radiology*. 2011;261:719–732.
7. Rao VM, Levin DC, Parker L, Cavanaugh B, Frangos AJ, Sunshine JH. How widely is computer-aided detection used in screening and diagnostic mammography? *J Am Coll Radiol*. 2010;7:802–805.
8. Malich A, Fischer DR, Böttcher J. CAD for mammography: The technique, results, current role and further developments. *Eur Radiol*. 2006;16:1449–1460.
9. Fenton JJ, Abraham L, Taplin SH, et al. Effectiveness of computer-aided detection in community mammography practice. *J Natl Cancer Inst*. 2011;103:1152–1161.
10. Lehman CD, Wellman RD, Buist DS, Kerlikowske K, Tosteson AN, Miglioretti DL, and Breast Cancer Surveillance Consortium. Diagnostic accuracy of digital screening mammography with and without computer-aided detection. *JAMA Intern Med*. 2015;175:1828–1837.
11. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521:436–444.
12. Schmidhuber J. Deep learning in neural networks: An overview. *Neural Netw*. 2015;61:85–117.
13. Mnih V, Kavukcuoglu K, Silver D, et al. Human-level control through deep reinforcement learning. *Nature*. 2015;518:529–533.
14. Silver D, Huang A, Maddison CJ, et al. Mastering the game of go with deep neural networks and tree search. *Nature*. 2016;529:484–489.
15. Cireşan DC, Giusti A, Gambardella LM, Schmidhuber J. Mitosis detection in breast cancer histology images with deep neural networks. In: *Medical Image Computing and Computer-Assisted Intervention*. vol 8150 of Lecture Notes in Computer Science. Berlin, Heidelberg: Springer; 2013:411–418.
16. Prasoon A, Petersen K, Igel C, Lauze F, Dam E, Nielsen M. Deep feature learning for knee cartilage segmentation using a triplanar convolutional neural network. *Med Image Comput Comput Assist Interv*. 2013;16:246–253.
17. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer Assisted Intervention*. Munich, Germany: Springer International Publishing; 2015;234–241.
18. Lake BM, Ullman TD, Tenenbaum JB, Gershman SJ. Building machines that learn and think like people. 2016. arXiv:1604.00289.
19. Ciompi F, De Hoop B, Van Riel SJ, et al. Automatic classification of pulmonary peri-fissural nodules in computed tomography using an ensemble of 2d views and a convolutional neural network out-of-the-box. *Med Image Anal*. 2015;26:195–202.
20. Bar Y, Diamant I, Wolf L, Greenspan H. Deep learning with non-medical training used for chest pathology identification. In: *Medical Imaging*. Proceedings of the SPIE, page 94140V. International Society for Optics and Photonics, 2015.
21. Hofmanninger J, Langs G. Mapping visual features to semantic profiles for retrieval in medical imaging. In: *Computer Vision and Pattern Recognition*. Boston, MA: IEEE Publishing; 2015;457–465.
22. Shin HC, Lu L, Kim L, Seff A, Yao J, Summers RM. Interleaved text/image deep mining on a very large-scale radiology database. In: *Computer Vision and Pattern Recognition*. Boston, MA: IEEE Publishing; 2015;1090–1099.
23. Shin HC, Roth HR, Gao M, et al. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Trans Med Imaging*. 2016;35:1285–1298.
24. Tajbakhsh N, Shin JY, Gurudu SR, et al. Convolutional neural networks for medical image analysis: Fine tuning or full training? *IEEE Trans Med Imaging*. 2016;35:1299–1312.
25. Russakovsky O, Deng J, Su H, et al. Imagenet large scale visual recognition challenge. *Int J Comput Vision*. 2014;115:1–42.
26. Oquab M, Bottou L, Laptev I, Sivic J. Learning and transferring midlevel image representations using convolutional neural networks. In: *Computer Vision and Pattern Recognition*. Columbus, OH: IEEE publishing; 2014;1717–1724 p.
27. Yue-Hei Ng J, Hausknecht M, Vijayanarasimhan S, Vinyals O, Monga R, Toderici G. Beyond short snippets: Deep networks for video classification. 2015. arXiv:150308909.
28. Karssemeijer N, te Brake G. Detection of stellate distortions in mammograms. *IEEE Trans Med Imaging*. 1996;15:611–619.
29. Kooi T, Karssemeijer N. Invariant features for discriminating cysts from solid lesions in mammography. In: *Breast Imaging*. Gifu: Springer, 2014:573–580.
30. Timp S, Karssemeijer N. A new 2D segmentation method based on dynamic programming applied to computer aided detection in mammography. *Med Phys*. 2004;31:958–971.
31. James B, Olivier B, Frédéric B, et al. Theano: A CPU and GPU math expression compiler. In: *Proceedings of the Python for Scientific Computing Conference (SciPy)*. vol 4, 2010;3.
32. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 2014. arXiv:14091556.
33. Kooi T, Litjens G, Van Ginneken B, et al. Large scale deep learning for computer aided detection of mammographic lesions. *Med Image Anal*. 2016;35:303–312.
34. Kooi T, Gubern-Mérida A, Mordang JJ, et al. A comparison between a deep convolutional neural network and radiologists for classifying regions of interest in mammography. In: Tingberg A, et al., editor, *Breast Imaging*. vol 9699 of Lecture Notes in Computer Science, Malmo: Springer International Publishing Switzerland, 2016;51–56.
35. Kingma D, Ba J. ADAM: A method for stochastic optimization. 2015. arXiv:14126980.

36. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A simple way to prevent neural networks from overfitting. *J Mach Learn Res*. 2014;15:1929–1958.
37. He K, Zhang X, Ren S, Sun J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *Computer Vision and Pattern Recognition*. 2015;1026–1034.
38. Friedman JH. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*. 2001:1189–1232.
39. Efron B. Bootstrap methods: Another look at the jackknife. *Ann Stat*. 1979;7:1–26.
40. Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. 2015. arXiv:14114038.
41. Bruna J, Mallat S. Invariant scattering convolution networks. *IEEE Trans Pattern Anal Mach Intell*. 2013;35:1872–1886.
42. Cohen TS, Welling M. Transformation properties of learned visual representations. 2014. arXiv:1412.7659.
43. Dieleman S, De Fauw J, Kavukcuoglu K. Exploiting cyclic symmetry in convolutional neural networks. 2016. arXiv:1602.02660.