Journal of Medical Imaging

MedicalImaging.SPIEDigitalLibrary.org

Classifying symmetrical differences and temporal change for the detection of malignant masses in mammography using deep neural networks

Thijs Kooi
Nico Karssemeijer

SPIE.

Thijs Kooi, Nico Karssemeijer, "Classifying symmetrical differences and temporal change for the detection of malignant masses in mammography using deep neural networks," *J. Med. Imag.* **4**(4), 044501 (2017), doi: 10.1117/1.JMI.4.4.044501.

Classifying symmetrical differences and temporal change for the detection of malignant masses in mammography using deep neural networks

Thijs Kooi* **and Nico Karssemeijer**
RadboudUMC Nijmegen, Diagnostic Image Analysis Group, Department of Radiology and Nuclear Medicine, Nijmegen, The Netherlands

Abstract. We investigate the addition of symmetry and temporal context information to a deep convolutional neural network (CNN) with the purpose of detecting malignant soft tissue lesions in mammography. We employ a simple linear mapping that takes the location of a mass candidate and maps it to either the contralateral or prior mammogram, and regions of interest (ROIs) are extracted around each location. Two different architectures are subsequently explored: (1) a fusion model employing two datastreams where both ROIs are fed to the network during training and testing and (2) a stagewise approach where a single ROI CNN is trained on the primary image and subsequently used as a feature extractor for both primary and contralateral or prior ROIs. A "shallow" gradient boosted tree classifier is then trained on the concatenation of these features and used to classify the joint representation. The baseline yielded an AUC of 0.87 with confidence interval [0.853, 0.893]. For the analysis of symmetrical differences, the first architecture where both primary and contralateral patches are presented during training obtained an AUC of 0.895 with confidence interval [0.877, 0.913], and the second architecture where a new classifier is retrained on the concatenation an AUC of 0.88 with confidence interval [0.859, 0.9]. We found a significant difference between the first architecture and the baseline at high specificity with $p = 0.02$. When using the same architectures to analyze temporal change, we yielded an AUC of 0.884 with confidence interval [0.865, 0.902] for the first architecture and an AUC of 0.879 with confidence interval [0.858, 0.898] in the second setting. Although improvements for temporal analysis were consistent, they were not found to be significant. The results show our proposed method is promising and we suspect performance can greatly be improved when more temporal data become available. [DOI: 10.1117/1.JMI.4.4.044501]

Keywords: deep learning; convolutional neural networks; machine learning; computer-aided diagnosis; breast cancer.

Paper 17072RR received Mar. 21, 2017; accepted for publication Sep. 12, 2017; published online Oct. 10, 2017.

1 Introduction

Breast cancer screening in the form of annual or biennial breast x-rays is being performed to detect cancer at an early stage. This has been shown to significantly increase chances of survival, with some studies showing a reduction in breast cancer mortality of up to 40%.[1] Human reading of screening data is time consuming and error prone; to aid interpretation, computer-aided detection and diagnosis (CAD)[2–5] systems are developed. For mammography, CAD is already widely applied as a second reader,[6,7] but the effectiveness of current technology is disputed. Several studies show no increase in sensitivity or specificity with CAD[8] for masses or even a decreased specificity without an improvement in detection rate or characterization of invasive cancers.[9,10]

During a mammographic exam, images are typically recorded of each breast, and the absence of a certain structure around the same location in the contralateral image will render an area under scrutiny more suspicious; conversely, the presence of a similar tissue less so. In addition, due to the annual or biennial organization of screening, there is a temporal dimension and similar principles apply: the amount of fibroglandular tissue is expected to decrease, rather than increase with age and, therefore, novel structures that are not visible on previous exams, commonly referred to as priors, spark suspicion.

In medical literature, an asymmetry denotes a potentially malignant density that is not characterized as a mass or architectural distortion. Four types are distinguished: (1) a plain asymmetry refers to a density lacking convex borders, seen in only one of the two standard mammographic views; (2) a focal asymmetry is visible on two views but does not fit the definition of a mass; (3) a global asymmetry indicates a substantial difference in total fibroglandular tissue between left and right breast; and (4) a developing asymmetry refers to a growing asymmetry in comparison to prior mammograms.[11,12] These types are generally benign but have been associated with an increased risk[13] and are sometimes the only manifestation of a malignancy. To the best of our knowledge, no relevant work has been done that compares reader performance of malignancies with and without left and right comparisons, but asymmetry is often mentioned by clinicians as an important clue also to detect malignancies that are classified as a mass. The merit of temporal comparison mammograms on the other hand has been well studied and is generally known to improve specificity without a profound impact on sensitivity for detection.[14–18]

Burnside et al.[15] analyzed a set of diagnostic and screening mammograms and concluded that in the latter case, comparison

*Address all correspondence to: Thijs Kooi, E-mail: email@thijskooi.com

2329-4302/2017/$25.00

with previous examinations significantly decreases the recall rate and false-positive rate, but does not increase sensitivity. Varela et al.[16] compared the reading performance of six readers and found the performance drops significantly when removing the prior mammogram, in particular in areas of high specificity, relevant for screening. Roelofs et al.[17] also investigated the merit of prior mammograms in both detection and assessment of malignant lesions. Their results show performance was significantly better in the presence of a prior exam, but no more lesions were found. They subsequently postulate priors are predominantly useful for interpretation and less so for initial detection. Yankaskas et al.[18] additionally investigated the effect of noticeable change in tissue in mammograms. They generated separate sets of current-prior examination pairs with and without noticeable change and observed that recall rate, sensitivity, and cancer detection rate are higher when change is noted but specificity is lower, resulting in a higher false-positive rate.

Symmetry is often used as a feature in traditional CAD systems detecting pathologies such as lesions in the brain,[19] prostate cancer,[20] and abnormalities in the lungs.[21] Most research on mammographic asymmetries involves the classification of a holistic notion of discrepancy rather than the incorporation of this information in a CAD system.[22,23] Published work on temporal analysis typically relies on the extraction of features from both current and prior exams, which are combined into a single observation and fed to a statistical learning algorithm.[24,25] For detection, an additional registration step is performed.[26] This has been shown to significantly increase the performance of the traditional, handcrafted feature-based systems.

Recent advances in machine learning, in particular deep learning,[27–30] signified a breakthrough in artificial intelligence and several pattern recognition applications are now claiming human or even super human performance.[31–34] Deep convolutional neural networks (CNNs)[27] are currently dominating leader boards in challenges for both natural[35] and medical image analysis challenges.[36–38] Rather than relying on engineers and domain experts to design features, the systems learn feature transformations from data, saving enormous amounts of time in development. The adoption of deep neural networks in medical image analysis was initially reluctant, but the community has recently seen a surge of papers[39] some showing significant improvements upon the state-of-the art.[40–43]

The vanilla CNN architecture is a generic problem solver for many signal processing tasks but is still limited by the constraint that a single tensor needs to be fed to the front-end layer, if no further adaptations to the network are made. Medical images provide an interesting new data source, warranting adaptation of methods successful in natural images. Several alternative architectures that go beyond the patch level and work with multiscale[44] or video[45–47] have been explored for natural scenes. In these settings, multiple datastreams are employed, where each datastream represents, for instance, a different scale in the image or frames at different time points in a video. Similar ideas have been applied to medical data, most notably the 2.5D simplification of volumetric scans.[40,48,49]

In this paper, we extend previous work[50] and investigate the addition of symmetry and temporal information to a deep CNN with the purpose of detecting malignant soft tissue lesions in mammography. We employ a simple linear mapping that takes the location of a mass candidate and maps it to either the contralateral or prior mammogram, and regions of interest (ROIs) are extracted around each location. We subsequently explore two different architectures:

1. A fusion model employing two datastreams where both ROIs are fed to the network during training and testing.
2. A stagewise approach where a single ROI CNN is trained on the primary image and subsequently used as feature extractor for both primary and contralateral or prior ROIs. A "shallow" gradient boosted tree (GBT) classifier is subsequently trained on the concatenation of these features and used to classify similar concatenations of features in the test set.

Examples of symmetry pairs are shown in Fig. 1. Figure 2 shows several examples of temporal pairs.

To the best of our knowledge, this is the first CAD and deep learning approach incorporating symmetry as a feature in a CAD system and the first CAD paper exploring deep neural networks for temporal comparison. Even though the methods are applied to mammography, we feel results may be relevant as well for other medical image analysis tasks, where the classification of anomalies that occur unilaterally or develop over time is important, such as lung, prostate, and brain images.

The rest of this paper is divided into five sections. In the following section, we will outline the data preprocessing, candidate detector, and linear mapping used. In Sec. 3, the deep neural architectures will be described followed by a description of

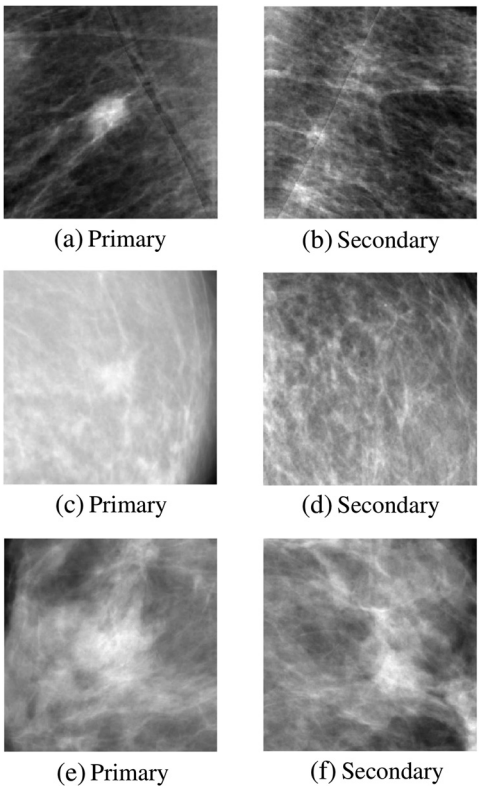


Fig. 1 Examples of symmetry pairs. Top row: Very suspicious malignant lesion (a) regardless of its contralateral counterpart (b). Middle row: Malignant lesion (a) that is more suspicious in the light of its contralateral image (b). Bottom row: Normal structure (a) that is less suspicious in the light of its contralateral image (b).

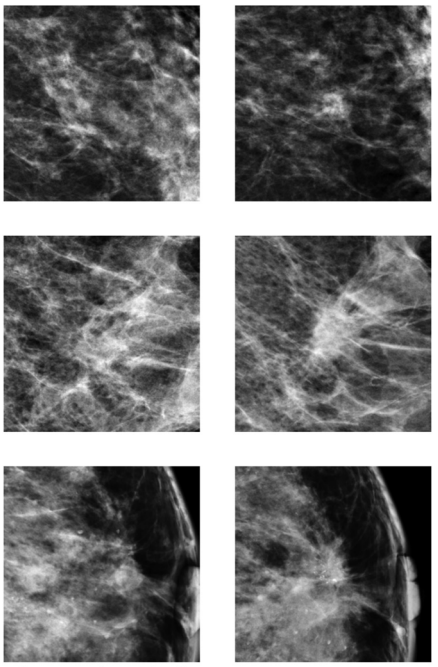


Fig. 2 Examples of temporal pairs. The right column represents the current and the left column the prior image it is compared with, using the mapping described in Sec. 2.2.

the data and experimental setup in Sec. 4. Results will be discussed in Sec. 5, and we will end with a conclusion in Sec. 6.

2 Methods

2.1 *Candidate Detection*

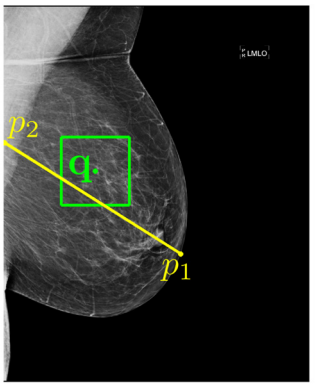
We generally follow the candidate detection setup described in Kooi et al.[41] To get potential locations of lesions and extract candidate patches, we make use of a popular candidate detector for mammographic lesions.[51] It employs five features based on first- and second-order Gaussian kernels, two designed to spot the center of a focal mass and two looking for spiculation patterns, characteristic of malignant lesions. A final feature indicates the size of optimal response in scale-space. We subsequently apply a random forest[52] classifier to generate a likelihood map on which we perform nonmaximum suppression. All optima are treated as candidates and patches of 250×250 pixels, or 5 cm at 200 μm, are extracted around each center location. Since many candidates are too close to the border to extract full patches, we pad the image with zeros.

For data augmentation, we follow the scheme described in Kooi et al.[41] Each patch in the training set containing an annotated malignant lesion is translated 16 times by adding values sampled uniformly from the interval $[-25, 25]$ (0.5 cm) to the lesion center. Each original positive patch is scaled 16 times by adding values sampled uniformly from the interval $[-30, 30]$ (0.6 cm) to the top left and bottom right of the bounding box. All patches, both positive and negative are rotated using four 90 deg rotations. This results in $(1 + 16 + 16)4 = 132$ patches per positive lesions and 4 per negative. In practice, these operations are computed on the fly during training, to prevent large datasets on disk. After candidates have been generated, locations are mapped to the same point in the contralateral image or the prior.

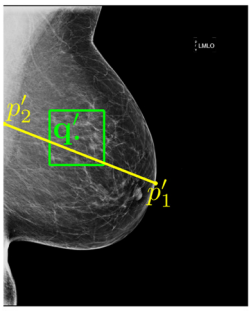
2.2 *Mapping Image Locations*

Finding corresponding locations between two mammograms is a challenging problem due to two main factors: (1) apart from the nipple and chest wall, which may not always be visible, there are no clear landmarks to accommodate feature-based registration and (2) the transformation is highly nonlinear. Before the mammogram is recorded, the breast is strongly deformed: the viewing area is optimized and dose is minimized by stretching the breast. In addition, the compression plates may not always touch the breast at the same location causing some movement of tissue within the breast.

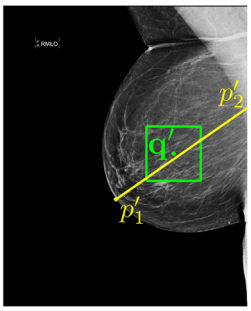
A comparative study among several commonly applied registration methods by Van Engeland et al.[53] found a simple linear approach based on the position of the nipple and center of mass alignment outperformed more complex methods such as warping. We propose a similar approach based on two landmarks. To obtain these points, the whole breast area is first segmented using simple thresholding, followed by a linear Hough transform to segment the pectoral muscle,[54] in the case of a mediolateral oblique (MLO) image. The row location of the front of the breast (an approximation of the nipple location) p_1 is subsequently estimated by taking a point on the contour of the breast with the largest distance to the line output by the Hough transform. A column point in the pectoral muscle or chest wall p_2 is taken by drawing a straight line from this



(a) Primary image



(b) Prior image



(c) Contralateral image

Fig. 3 To incorporate symmetry and temporal information, we make use of a simple mapping, based on two coordinates indicated by the end points of the yellow line. (a) A region of interest (ROI) represented by the green box is extracted around a potential malignant lesion location, indicated by the green dot, found by a candidate detector. The location is subsequently matched to either (b) the prior or (c) the contralateral image. We explore two deep convolutional neural network (CNN) fusion strategies to optimally capture the relation between contralateral and prior images.

point perpendicular to the fit output by the Hough transform. The lesion center in the image under evaluation $\mathbf{q} = (q_r, q_c)^T$, where q_r and q_c denote the row and column location, respectively, is subsequently mapped to the estimated lesion center $\mathbf{q}'$ in the contralateral or prior image according to

$$\mathbf{q}' = \mathbf{q} - \mathbf{p} + \mathbf{p}', \tag{1}$$

with $\mathbf{p} = (p_1, p_2)^T$ and $\mathbf{p}' = (p_1', p_2')^T$ the same points in the contralateral or prior mammogram. In other words, we simply clamp the x-distance to the chest wall and the y-distance to the estimated location of the nipple. An example is provided in Fig. 3.

Since most CNN architectures induce a decent amount of translation invariance, the mapping does not need to be very precise. To further mitigate mapping errors, we introduce a form of data augmentation by mapping each location in the image in question to 64 different points in the comparison mammogram by sampling the location from a Gaussian with zero mean and 10 pixel standard deviation.

3 Deep Convolutional Neural Networks

The CNN architecture exploits structure of the input by sharing weights at different locations in the image, resulting in a convolution operation, the main workhorse of the CNN. The main difference between deep models and conventional statistical learning methods is the nested nonlinear function the architecture represents. At each layer, the input signal is convolved with a set of K kernels $\mathcal{W} = (\mathbf{W}_1, \mathbf{W}_2, \ldots, \mathbf{W}_K)$ and biases $\mathcal{B} = \{b_1, b_2, \ldots, b_K\}$ are added, each generating a new set of feature maps $\mathbf{X}_k$. These features are subjected to an elementwise nonlinear transform $\sigma(\cdot)$ and the same process is repeated for every convolutional layer $l_0, l_1, \ldots, l_L$

$$\mathbf{X}_k^l = \sigma(\mathbf{W}_k^{l-1} \otimes \mathbf{X}^{l-1} + b_k^{l-1}). \tag{2}$$

Convolutional layers are generally alternated with pooling layers that subsample the resulting feature maps, generating some translation invariance and reducing the dimensionality as information flows through the architecture. After these layers, the final tensor of feature maps is flattened to a vector $\mathbf{x}^l$ and several fully connected layers are typically added, where weights are no longer shared

$$\mathbf{x}^l = \sigma(\mathbf{W}^l \mathbf{x}^{l-1} + b^l). \tag{3}$$

The posterior distribution over a class variable y_i, given input patch $\mathbf{X}^0$ is acquired by feeding the last level of activations $\mathbf{x}^L$ to either a logistic sigmoid for single class or a softmax function for multiclass

$$P(y_i|\mathbf{X}^0;\Theta) = \text{softmax}(\mathbf{x}^L;\mathbf{W},\mathbf{b}) = \frac{e^{\mathbf{w}_i^T \mathbf{x}^L + b_i^L}}{\sum_{k=1}^K e^{\mathbf{w}_k^T \mathbf{x}^L + b_k^L}}, \tag{4}$$

with Θ is the set of all weights and biases in the network and $\mathbf{w}_i$ is the vectorized set of weights leading to the output node of class i. The whole network can be seen as a parameterized feature extractor and classifier, where the parameters of the feature transformation and classifier are learned jointly and optimized based on training data.

The parameters in the network are generally learned using maximum likelihood estimation or maximum a-posteriori, when employing regularization and default backpropagation. Increasing depth up to some point seems to improve efficiency and reduce the amount of parameters that need to be learned, without sacrificing performance or even increases overall performance.[55–57] The gradient of the error of each training sample is dispersed among parameters in every layer during backpropagation and hence becomes smaller (or in rare cases explodes), which is referred to as the fading gradient problem. Common tricks to quell this phenomenon are smart weight initialization,[58,59] batch normalization,[33] and nonsaturating transfer functions such as rectified linear units or recently exponential linear units (ELU).[60]

3.1 *Fusion Architectures*

Partly inspired by the work of Karpathy et al.,[45] we propose to add the contralateral and (first prior) temporal counterparts of a patch as separate datastreams to a network. In principle, the datastreams can be merged at any point in the network, with simply treating the additional patch as a second channel the extreme case. Neverova et al.[46] postulated the optimal point of fusion pertains to the degree of similarity of the sources, but to the best of our knowledge, no empirical or theoretical work exists that investigates this. We evaluate two architectures:

1. A two-stream network where kernels are shared and datastreams are fused at the first fully connected layer. Figure 4 provides an illustration of this network.
2. A single patch, single stream network is used as a feature extractor by classifying all samples in the training and test set and extracting the latent representation of each patch from the first fully connected layer $\mathbf{x}^{fc1}$ of the network. This feature representation of the primary and either contralateral or prior ROI are concatenated and fed to a "shallow" GBT classifier to generate a new posterior that captures both symmetry or (first prior) temporal information.

The second approach is far easier to train, since it does not entail reoptimizing hyperparameters of a deep model, which is tedious and time consuming. A downside is that the kernels effectively see less data and are therefore potentially less optimal for the task. In addition, the second setup is more prone to overfitting. We will elaborate on this in the discussion.

In general, there are a lot less temporal than symmetry samples because they require two rounds of screening and symmetry samples only one. To compare these architectures, we could simply take a subset of the data where each current exam has both a contralateral and prior counterpart. Unfortunately, this yields a relatively small number of positive samples, and in early experiments, we found the (base) performance to be very marginal and not sufficient to provide a fair comparison. We therefore view missing prior exams simply as missing data. Although missing data have been well studied in the statistics community,[61] relatively little has been published with respect to discriminative models.

In the context of recurrent neural networks (RNNs),[62–64] several imputation methods have been explored.[65,66] Lipton et al.[66] investigated two imputation strategies: zero-imputation, where missing samples are simply set to zero and forward-filling that sets the missing value to the value observed before that. Their results show zero imputation with missing data indicators

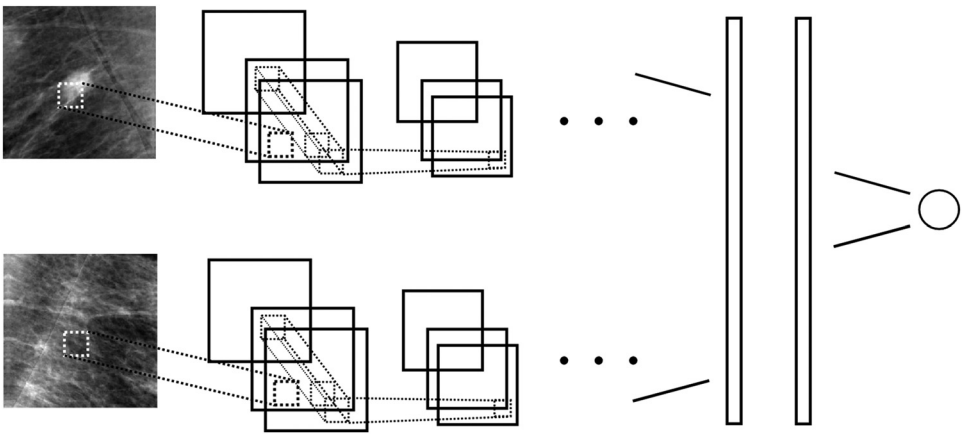


Fig. 4 To learn differences between left and right breast and temporal change around a candidate location, we use a two-stream convolutional neural network (CNN). The first stream has as input a patch centered at a candidate location, the second stream a patch around the same location in either the contralateral image or the prior, using the mapping depicted in Fig. 3. All weights are shared across streams and feature maps are concatenated before the first fully connected layer.

works best, but no significance analysis is performed. In a similar spirit, we explore two strategies:

1. Use a black image when no prior is available. When a woman skipped a screening round, we map the image to the exam 4 years before the current or add a black image if this is absent.
2. Use the image from the exam 4 years before the current image and use the current when no prior is available.

The first approach carries some additional information, in the sense that the absence of a prior may also increase the likelihood that an exam is positive, since more cancers are typically found in the first round of screening. In the second setting, it is difficult for the network to distinguish pairs where no change is observed and pairs where simply no prior is available.

To add symmetry and temporal information simultaneously, both architectures can trivially be extended with a third stream. However, this requires some additional engineering; we, therefore, restrict this study to learning two separate models and will propose ways to extend this in the discussion.

4 Experiments

4.1 *Data*

Our data were collected from a mammography screening program in the Netherlands (screening midwest) and was recorded with a Hologic Selenia mammography device at an original resolution of 70 μm. All malignant masses were biopsy proven and annotated using contours drawn under the supervision of experienced radiologists. A candidate was considered positive, if the locations were in or within 0.7 cm from an annotated malignant lesion. Before presentation to the human reader, the image is typically processed to optimize contrast and enhance the breast periphery. To prevent information loss, we work on the raw images instead and only apply a log transform that results in a representation in which attenuation and pixel values are linearly related. The images are subsequently scaled to 200 μm using bilinear interpolation.

Table 1 Overview of the data used for training, validation, and testing. Findings refer to the amount of candidates (before data augmentation). Number are separated by "/" where the first number indicates the amount for training, the second the amount for validation, and the third the amount for testing.

	Findings	Cases
Masses	869/210/470	796/189/386
Normal	200,982/54,566/74,799	3111/1482/1137

Our dataset consists of 18,366 cases of 18,366 women. Each case comprises one or more exams taken at intervals of 2 years, unless a women skipped a screening. Each exam again typically consists of four images: a craniocaudal and MLO view of each breast. We generated training, validation, and test set by splitting on a case level, i.e., samples from the same patient are not scattered across sets. We took 65% for training, 15% for validation, and 25% for testing. Annotated benign samples were removed from our training set but kept in the test set. Since not all benign samples are annotated in our dataset, we cannot provide reasonable estimates of the amount of samples. An overview of the data is provided in Table 1.

4.2 *Learning Settings and Implementation Details*

The networks were implemented in TensorFlow[67] and generally follow the architecture used in Kooi et al.[41] Hyperparameters of all models were optimized on a separate validation set using random search.[68] For the deep CNNs, we employed VGG-like[56] architectures with five convolutional layers with {16, 16, 32, 32, 64} kernels of size 3×3 in all layers. We used "valid" convolutions using a stride of 1 in all settings. Max pooling of 2×2 was used, using a stride of 1 in all but the final convolutional layer. Two fully connected layers of 512 each were added. Weights were initialized using the MSRA weight filler,[59] with weight sampled from a truncated normal, all biases were initialized to 0.001. We employed ELU's[60] as transfer functions in all

layers. Learning rate, dropout rate, and L2-norm coefficient were optimized per architecture.

Since the class ratio is in the order of 1/10,000, randomly sampling minibatches will result in very poor performance as the network will just learn to classify all samples as negative. We, therefore, applied the following scheme. We generated two separate datasets, one for all positive and one for all negative samples. Negative samples are read from the disk in chunks and all positive samples are loaded into host RAM. During an epoch, we cycle through all negative samples and in each minibatch take a random selection of an equal amount of positives, which are subsequently fed to GPU where gradients are computed and updated. This way, all negative samples are presented in each epoch, and the class balance is maintained.

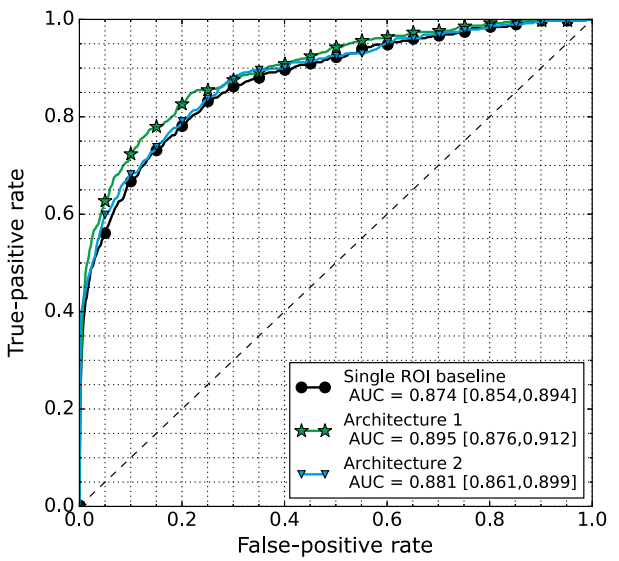


Fig. 5 ROC curves of the baseline CNN using a single ROI and the two fusing architectures described in Sec. 3.1 when presented with the contralateral ROI.

Each configuration trained for roughly 10 days on a TitanX 12 GB GPU.

For the shallow model, we employ GBTs[69] using the excellent XGBoost implementation.[70] We cross validated the shrinkage and depth using 16 folds. Further parameters were tuned on a fixed validation set using a coordinate descent such as scheme. Since the last fully connected layer has size 512, the input to the GBT comprised of 512 features for the single patch setting and a feature vector of 1024 in the symmetry and temporal setting.

4.3 *Results*

Given the results from clinical literature regarding the merit of priors, we focus our results on the classification of candidates and, therefore, only present ROC curves, rather than FROC curves that are commonly used for detection. To obtain confidence intervals and perform significance testing, we performed bootstrapping[71] using 5000 bootstraps. All curves shown are the mean curve from these bootstrap samples using cubic interpolation. The baseline yielded an AUC of 0.870 with confidence interval [0.853, 0.893].

Figure 5 shows the results of the single ROI baseline, and the fusion architectures as described in Sec. 3.1 applied to the symmetry comparison. The first architecture where both patches are presented during training yielded an AUC of 0.895 with confidence interval [0.877, 0.913] and the second architecture where a new classifier is retrained on the concatenation yielded an AUC of 0.880 with confidence interval [0.859, 0.900]. We find significant difference at high specificity on the interval $[0, 0.2]$, $p = 0.02$ between the first architecture and the baseline, but no significant difference on the full AUC ($p = 0.14$). For the second architecture, we did not find a significant difference between either the baseline or the first architecture.

Figure 6 shows the results of the single ROI baseline and the fusion architectures applied to the temporal comparison. We first investigated the difference between the two different strategies to handle missing priors. The approach using the same image obtained an AUC of 0.873 with confidence interval [0.854, 0.892], the approach using the black image for missing priors an AUC of 0.884 with confidence interval [0.866, 0.902]. We

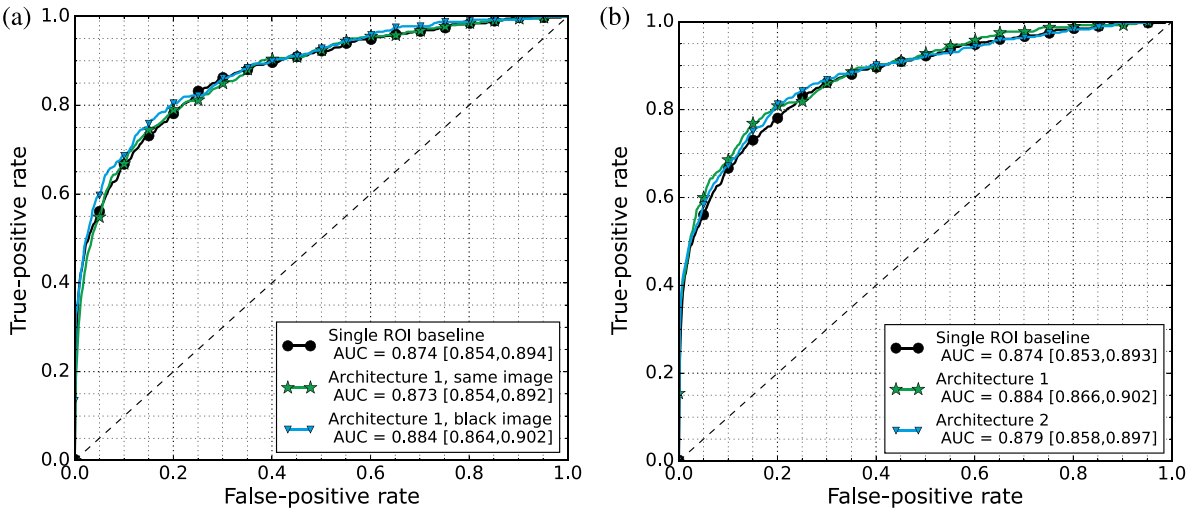


Fig. 6 (a) ROC curves of the baseline CNN using a single ROI and the two strategies to handle missing prior images both using architecture 1. (b) ROC curves of the baseline CNN using a single ROI and the two fusing architectures described in Sec. 3.1 when presented with the prior ROI and black image strategy.

did not see a significant difference among the strategies $p \gg 0.05$; however, the strategy where the black image was used has a higher AUC, and we have decided to use this to compare the fusing architectures.

The first architecture where both patches are presented during training obtained an AUC of 0.884 with confidence interval [0.866, 0.902] and the second architecture where a new classifier is retrained on the concatenation of both primary and symmetry or temporal features reached an AUC of 0.879 with confidence interval [0.858, 0.898]. We did not find a significant difference among any of the architectures $p \gg 0.05$, but improvements were found to be consistent during early experiments. Results will be discussed in the following section.

5 Discussion

From the curves in Figs. 5 and 6, we can see both symmetry and temporal data improve performance but only see marginal improvements with temporal data. The curves also show the scheme where both ROIs are fed to a single network [architecture (1) in Sec. 3.1] work best. As mentioned in Sec. 3.1, architecture (2) has the advantage that no new networks need to be trained, which can take several months to do properly for large datasets. Two disadvantages, however, are that (1) the kernels in the network (parameters up to the first fully connected layer) effectively see less data. In the first architecture, even though the kernels are shared, they are trained on both the primary and either symmetry or prior patch and, therefore, better adjusted to the task. (2) Overfitting is a much bigger issue: since the features are learned on most of the data the models are trained on, the cross validation procedure of the GBT often gave a strong underestimate of the optimal regularization coefficients (depth, shrinkage in the case of the GBT), resulting in strong gaps between train and test performance. Optimizing this on a fixed validation set did not result in much better performance. We have tried extracting features from deeper in the network to mitigate this effect but found lower performance.

Since many exams do not have a prior, we explored two strategies to fill in this missing data. In the first setting, we used a black image when no prior image was available and in the second strategy, the same image as the current was used. From the curves in Fig. 6, we can see that in the first setting the prior ROI does add some information; therefore, this approach is at least not detrimental to performance. In the second setting, however, we do not see an increase. A possible advantage of the first approach is that it carries some additional information: the number of tumors found in the first screening round is often higher; when using imputation methods mentioned by Lipton et al.[64] this information is effectively lost. As also mentioned in Sec. 3.1, the disadvantage of the second approach is that it is difficult for the network to distinguish between malignant mass-no prior pairs and malignant mass-malignant mass pairs, since no change is typically associated with normal tissue.

In clinical practice, radiologists sometimes look back two studies instead of one, when comparing the current to the prior. Since this requires three screening rounds, this reduces the size of our dataset again, if we want to emulate this and more prior ROIs need an imputed image. Ideally, the neural network architecture should accommodate a varying set of priors. In early experiments, we have explored the use of RNNs,[62–64] a model designed for temporal data that can be trained and tested on varying input and output sizes. We did not see a clear improvement in performance but plan to explore this idea more in future work. Since this model can work with varying length inputs, it also provides an elegant way to handle missing prior exams.

In this study, we have trained all networks from scratch. Since the rudimentary features that are useful to detect cancer in one view are expected to be almost as useful when combining views, a better strategy may be to initialize the symmetry or temporal two-stream network with the weights trained on a single ROI. Similarly, since we expect similar features are useful to spot discrepancies between left and right breast as to spot differences between time points, the temporal network could be initialized with the network trained on symmetry patches or the other way around. Due to time constraints, this was left to future work, but we suspect an increase in performance.

We have compared two different fusion strategies. As mentioned in Sec. 3.1, the datastreams can in principle be fused at any point in the network, as done by Karpathy et al.[45] However, there is no guarantee that different architectures perform optimal using the same hyperparameters. For instance, the weight updates of lower layers change if fusion is performed at different points higher in the network. In particular, the learning rate is often found to be important and we feel comparison rings somewhat hollow if no extensive search through the parameter space is done. Since a model typically trains for roughly a week, this is infeasible with our current hardware and we have decided to focus on the two presented models.

Since the focus of this paper is the presentation of two fusion schemes for adding symmetry and temporal information to a deep CNN, we have presented separate results for each. In practice, when using a CAD system to generate a label for a case, these should be merged into one decision. As mentioned in Sec. 3.1, extending the network with a third datastream is trivial. However, this limits the application to cases where both prior and contralateral image are available. In our method, we have added a black image, where priors were not available, and a similar approach could be pursued in this setting. Another option would be to train a third classifier on top of the latent representation from separate CNNs or the posterior output by separate CNNs, possibly using a missing data model. Since training deep neural networks and optimizing hyperparameters takes a lot of time, we have left this for future work.

6 Conclusion

In this paper, we have presented two deep CNN architectures to add symmetry and temporal information to a computer-aided detection (CAD) system for mass candidates in mammography. To the best of our knowledge, this is the first approach exploring deep CNNs for symmetry and temporal classification in a CAD system. Results show improvement in performance for both symmetry and temporal data. Though in the latter case gain in performance is still marginal, it is promising and we suspect that when more data become available, performance will significantly increase. Although the methods are applied to mammography, we think results can be relevant for other CAD problems where symmetrical differences within or between organs are sought, such as lung, brain, and prostate images or CAD tasks where temporal change needs to be analyzed, such as lung cancer screening.

Disclosures

Thijs Kooi has no potential conflicts of interest. Nico Karssemeijer is co-founder, shareholder, and director of

ScreenPoint Medical BV (Nijmegen, The Netherlands), co-founder of Volpara Health Technologies Ltd. (Wellington, New Zealand), and QView Medical Inc. (Los Altos, California).

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.

References

1. L. Tabar et al., "Mammography service screening and mortality in breast cancer patients: 20-year follow-up before and after introduction of screening," *Lancet* **361**, 1405–1410 (2003).
2. M. L. Giger, N. Karssemeijer, and S. G. Armato, "Computer-aided diagnosis in medical imaging," *IEEE Trans. Med. Imaging* **20**, 1205–1208 (2001).
3. K. Doi, "Computer-aided diagnosis in medical imaging: historical review, current status and future potential," *Comput. Med. Imaging Graph.* **31**, 198–211 (2007).
4. K. Doi, "Current status and future potential of computer-aided diagnosis in medical imaging," *Br. J. Radiol.* **78**(Suppl. 1), S3–S19 (2005).
5. B. van Ginneken, C. M. Schaefer-Prokop, and M. Prokop, "Computer-aided diagnosis: how to move from the laboratory to the clinic," *Radiology* **261**(3), 719–732 (2011).
6. V. M. Rao et al., "How widely is computer-aided detection used in screening and diagnostic mammography?" *J. Am. Coll. Radiol.* **7**, 802–805 (2010).
7. A. Malich, D. R. Fischer, and J. Böttcher, "CAD for mammography: the technique, results, current role and further developments," *Eur. Radiol.* **16**, 1449–1460 (2006).
8. P. Taylor et al., "Impact of computer-aided detection prompts on the sensitivity and specificity of screening mammography," *Health Technol. Assess.* **9**, iii 1–58 (2005).
9. J. J. Fenton et al., "Effectiveness of computer-aided detection in community mammography practice," *J. Natl. Cancer Inst.* **103**, 1152–1161 (2011).
10. C. D. Lehman et al., "Diagnostic accuracy of digital screening mammography with and without computer-aided detection," *JAMA Intern. Med.* **175**, 1828–1837 (2015).
11. E. A. Sickles, "The spectrum of breast asymmetries: imaging features, work-up, management," *Radiol. Clin.* **45**(5), 765–771 (2007).
12. J. H. Youk et al., "Asymmetric mammographic findings based on the fourth edition of BI-RADS: types, evaluation, and management," *Radiographics* **29**(1), e33 (2009).
13. D. Scutt, G. A. Lancaster, and J. T. Manning, "Breast asymmetry and predisposition to breast cancer," *Breast Cancer Res.* **8**(2), R14 (2006).
14. M. G. Thurfjell et al., "Effect on sensitivity and specificity of mammography screening with or without comparison of old mammograms," *Acta Radiol.* **41**(1), 52–56 (2000).
15. E. S. Burnside et al., "Differential value of comparison with previous examinations in diagnostic versus screening mammography," *Am. J. Roentgenol.* **179**(5), 1173–1177 (2002).
16. C. Varela et al., "Use of prior mammograms in the classification of benign and malignant masses," *Eur. J. Radiol.* **56**, 248–255 (2005).
17. A. A. J. Roelofs et al., "Importance of comparison of current and prior mammograms in breast cancer screening," *Radiology* **242**, 70–77 (2007).
18. B. C. Yankaskas et al., "Effect of observing change from comparison mammograms on performance of screening mammography in a large community-based population," *Radiology* **261**(3), 762–770 (2011).
19. S. X. Liu, "Symmetry and asymmetry analysis and its implications to computer-aided diagnosis: a review of the literature," *J. Biomed. Inf.* **42**(6), 1056–1064 (2009).
20. G. Litjens et al., "Computer-aided detection of prostate cancer in MRI," *IEEE Trans. Med. Imaging* **33**, 1083–1092 (2014).
21. B. van Ginneken, "Computer-aided diagnosis in chest radiography," PhD Thesis, Utrecht University, The Netherlands (2001).
22. R. J. Ferrari et al., "Analysis of asymmetry in mammograms via directional filtering with Gabor wavelets," *IEEE Trans. Med. Imaging* **20**, 953–964 (2001).
23. P. Casti et al., "Analysis of structural similarity in mammograms for detection of bilateral asymmetry," *IEEE Trans. Med. Imaging* **34**(2), 662–671 (2015).
24. L. Hadjiiski et al., "Analysis of temporal changes of mammographic features: computer-aided classification of malignant and benign breast masses," *Med. Phys.* **28**, 2309–2317 (2001).
25. S. Timp, C. Varela, and N. Karssemeijer, "Temporal change analysis for characterization of mass lesions in mammography," *IEEE Trans. Med. Imaging* **26**, 945–953 (2007).
26. S. Timp and N. Karssemeijer, "Interval change analysis to improve computer aided detection in mammography," *Med. Image Anal.* **10**, 82–95 (2006).
27. Y. Lecun et al., "Gradient-based learning applied to document recognition," *Proc. IEEE* **86**, 2278–2324 (1998).
28. G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.* **18**, 1527–1554 (2006).
29. Y. Bengio, A. Courville, and P. Vincent, "Representation learning: a review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.* **35**(8), 1798–1828 (2013).
30. J. Schmidhuber, "Deep learning in neural networks: an overview," *Neural Networks* **61**, 85–117 (2015).
31. D. C. Cireşan et al., "Multi-column deep neural network for traffic sign classification," *Neural Networks* **32**, 333–338 (2012).
32. V. Mnih et al., "Human-level control through deep reinforcement learning," *Nature* **518**, 529–533 (2015).
33. S. Ioffe and C. Szegedy, "Batch normalization: accelerating deep network training by reducing internal covariate shift," in *Int. Conf. on Machine Learning*, pp. 448–456 (2015).
34. D. Silver et al., "Mastering the game of Go with deep neural networks and tree search," *Nature* **529**, 484–489 (2016).
35. O. Russakovsky et al., "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vision* **115**(3), 211–252 (2014).
36. D. C. Cireşan et al., "Mitosis detection in breast cancer histology images with deep neural networks," *Lect. Notes Comput. Sci.* **8150**, 411–418 (2013).
37. O. Ronneberger, P. Fischer, and T. Brox, "U-Net: convolutional networks for biomedical image segmentation," *Lect. Notes Comput. Sci.* **9351**, 234–241 (2015).
38. D. Wang et al., "Deep learning for identifying metastatic breast cancer," arXiv preprint arXiv:1606.05718 (2016).
39. G. Litjens et al., "A survey on deep learning in medical image analysis," *Med. Image Anal.* **42**, 60–88 (2017).
40. H. Roth et al., "Improving computer-aided detection using convolutional neural networks and random view aggregation," *IEEE Trans. Med. Imaging* **35**, 1170–1181 (2016).
41. T. Kooi et al., "Large scale deep learning for computer aided detection of mammographic lesions," *Med. Image Anal.* **35**, 303–312 (2017).
42. A. A. A. Setio et al., "Pulmonary nodule detection in CT images: false positive reduction using multi-view convolutional networks," *IEEE Trans. Med. Imaging* **35**(5), 1160–1169 (2016).
43. M. J. J. P. van Grinsven et al., "Fast convolutional neural network training using selective data sampling: application to hemorrhage detection in color fundus images," *IEEE Trans. Med. Imaging* **35**(5), 1273–1284 (2016).
44. C. Farabet et al., "Learning hierarchical features for scene labeling," *IEEE Trans. Pattern Anal. Mach. Intell.* **35**(8), 1915–1929 (2013).
45. A. Karpathy et al., "Large-scale video classification with convolutional neural networks," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 1725–1732, IEEE (2014).
46. N. Neverova et al., "Multi-scale deep learning for gesture detection and localization," in *European Conf. on Computer Vision* (2014).
47. K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Int. Conf. on Neural Information Processing Systems* (2014).
48. A. Prasoon et al., "Deep feature learning for knee cartilage segmentation using a triplanar convolutional neural network," *Lect. Notes Comput. Sci.* **8150**, 246–253 (2013).
49. H. R. Roth et al., "A new 2.5D representation for lymph node detection using random sets of deep convolutional neural network observations," *Lect. Notes Comput. Sci.* **8673**, 520–527 (2014).
50. T. Kooi and N. Karssemeijer, "Deep learning of symmetrical discrepancies for computer-aided detection of mammographic masses," *Proc. SPIE* **10134**, 101341J (2017).

51. N. Karssemeijer and G. te Brake, "Detection of stellate distortions in mammograms," *IEEE Trans. Med. Imaging* **15**, 611–619 (1996).
52. L. Breiman, "Random forests," *Mach. Learn.* **45**(1), 5–32 (2001).
53. S. van Engeland et al., "A comparison of methods for mammogram registration," *IEEE Trans. Med. Imaging* **22**, 1436–1444 (2003).
54. N. Karssemeijer, "Automated classification of parenchymal patterns in mammograms," *Phys. Med. Biol.* **43**, 365–378 (1998).
55. R. K. Srivastava, K. Greff, and J. Schmidhuber, "Training very deep networks," in *Advances in Neural Information Processing Systems*, pp. 2377–2385 (2015).
56. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556 (2014).
57. K. He et al., "Deep residual learning for image recognition," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778 (2016).
58. X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Int. Conf. on Artificial Intelligence and Statistics*, pp. 249–256 (2010).
59. K. He et al., "Delving deep into rectifiers: surpassing human-level performance on ImageNet classification," in *IEEE Int. Conf. on Computer Vision and Pattern Recognition*, pp. 1026–1034 (2015).
60. D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (ELUs)" (2015).
61. P. D. Allison, *Missing Data*, Sage Publications, Thousand Oaks, California (2001).
62. A. Graves, *Supervised Sequence Labelling wth Recurrent Neural Networks*, Vol. **385**, Springer, Heidelberg (2012).
63. K. Greff et al., "LSTM: a search space odyssey," *IEEE Trans. Neural Networks Learn. Syst.* **PP**, 1–11 (2016).
64. Z. C. Lipton, J. Berkowitz, and C. Elkan, "A critical review of recurrent neural networks for sequence learning," arXiv preprint arXiv:1506.00019 (2015).
65. Z. Che et al., "Recurrent neural networks for multivariate time series with missing values," arXiv preprint arXiv:1606.01865 (2016).
66. Z. C. Lipton, D. C. Kale, and R. Wetzel, "Modeling missing data in clinical time series with RNNs," in *Machine Learning for Healthcare* (2016).
67. M. Abadi et al., "Tensorflow: large-scale machine learning on heterogeneous distributed systems," arXiv:1603.04467 (2016).
68. J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *J. Mach. Learn. Res.* **13**(1), 281–305 (2012).
69. J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Ann. Stat.* **29**, 1189–1232 (2001).
70. T. Chen and C. Guestrin, "XGBoost: a scalable tree boosting system," in *Proc. of the 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, pp. 785–794, ACM (2016).
71. B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*, Vol. **57**, CRC Press, Boca Raton (1994).

Thijs Kooi is a PhD candidate at the Diagnostic Image Analysis Group of the RadboudUMC Nijmegen, working on computer-aided diagnosis of breast cancer. He obtained his MSc degree in artificial intelligence from the University of Amsterdam in 2012 (cum laude) and his BSc degree in the same field from the University of Groningen. He held visiting research positions at Keio University in Japan, the National University of Singapore, and Johns Hopkins University. His main research interests are machine learning applied to medical image analysis and higher level decision making.

Nico Karssemeijer is a professor of computer-aided diagnosis. He studied physics at Delft University of Technology and graduated from Radboud University Nijmegen, Department of Medical Physics. In 1989, he joined the Department of Radiology, Radboud University Nijmegen Medical Center, where he formed a research group in computer aided detection (CAD). His professorship is in the Faculty of Science of the Radboud University in the section Intelligent Systems of the Institute for Computing and Information Sciences iCIS.