Medical Image Analysis 35 (2017) 303–312

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Large scale deep learning for computer aided detection of mammographic lesions

Thijs Kooi[a,*], Geert Litjens[a], Bram van Ginneken[a], Albert Gubern-Mérida[a], Clara I. Sánchez[a], Ritse Mann[a], Ard den Heeten[b], Nico Karssemeijer[a]

[a] Diagnostic Image Analysis Group, Department of Radiology, Radboud University Medical Center, Nijmegen, The Netherlands
[b] Department of Radiology, University Medical Centre Amsterdam, Amsterdam, The Netherlands

ARTICLE INFO

Article history:
Received 11 February 2016
Revised 12 July 2016
Accepted 20 July 2016
Available online 2 August 2016

Keywords:
Computer aided detection
Mammography
Deep learning
Machine learning
Breast cancer
Convolutional neural networks

ABSTRACT

Recent advances in machine learning yielded new techniques to train deep neural networks, which resulted in highly successful applications in many pattern recognition tasks such as object detection and speech recognition. In this paper we provide a head-to-head comparison between a state-of-the art in mammography CAD system, relying on a manually designed feature set and a Convolutional Neural Network (CNN), aiming for a system that can ultimately read mammograms independently. Both systems are trained on a large data set of around 45,000 images and results show the CNN outperforms the traditional CAD system at low sensitivity and performs comparable at high sensitivity. We subsequently investigate to what extent features such as location and patient information and commonly used manual features can still complement the network and see improvements at high specificity over the CNN especially with location and context features, which contain information not available to the CNN. Additionally, a reader study was performed, where the network was compared to certified screening radiologists on a patch level and we found no significant difference between the network and the readers.

1. Introduction

Nearly 40 million mammographic exams are performed in the US alone on a yearly basis, arising predominantly from screening programs implemented to detect breast cancer at an early stage, which has been shown to increase chances of survival (Tabar et al., 2003; Broeders et al., 2012). Similar programs have been implemented in many western countries. All this data has to be inspected for signs of cancer by one or more experienced readers which is a time consuming, costly and most importantly error prone endeavor. Striving for optimal health care, Computer Aided Detection and Diagnosis (CAD) (Giger et al., 2001; Doi, 2007; 2005; van Ginneken et al., 2011) systems are being developed and are currently widely employed as a second reader (Rao et al., 2010; Malich et al., 2006), with numbers from the US going up to 70% of all screening studies in hospital facilities and 85% in private institutions (Rao et al., 2010). Computers do not suffer from drops in concentration, are consistent when presented with the same input data and can potentially be trained with an incredible amount of training samples, vastly more than any radiologist will experience in his lifetime.

Until recently, the effectiveness of CAD systems and many other pattern recognition applications depended on meticulously handcrafted features, topped off with a learning algorithm to map it to a decision variable. Radiologists are often consulted in the process of feature design and features such as the contrast of the lesion, spiculation patterns and the sharpness of the border are used, in the case of mammography. These feature transformations provide a platform to instill task-specific, a-priori knowledge, but cause a large bias towards how we humans think the task is performed. Since the inception of Artificial Intelligence (AI) as a scientific discipline, research has seen a shift from rule-based, problem specific solutions to increasingly generic, problem agnostic methods based on learning, of which *deep learning* (Bengio, 2009; Bengio et al., 2013; Schmidhuber, 2015; LeCun et al., 2015) is its most recent manifestation. Directly distilling information from training samples, rather than the domain expert, deep learning allows us to optimally exploit the ever increasing amounts of data and reduce human bias. For many pattern recognition tasks, this has proven to be successful to such an extent that systems are now reaching human or even superhuman performance (Cireşan et al., 2012; Mnih et al., 2015; He et al., 2015).

* Corresponding author.
E-mail address: thijs.kooi@radboudumc.nl, email@thijskooi.com (T. Kooi).

http://dx.doi.org/10.1016/j.media.2016.07.007
1361-8415/

The term *deep* typically refers to the layered non-linearities in the learning systems, which enables the model to represent a function with far less parameters and facilitates more efficient learning (Bengio et al., 2007; Bengio, 2009). These models are not new and work has been done since the late seventies (Fukushima, 1980; Lecun et al., 1998). In 2006, however, two papers (Hinton et al., 2006; Bengio et al., 2007) showing deep networks can be trained in a greedy, layer-wise fashion sparked new interest in the topic. Restricted Boltzmann Machines (RBM's), probabilistic generative models, and autoencoders (AE), one layer neural networks, were shown to be expedient pattern recognizers when stacked to form Deep Belief Networks (DBN) (Hinton et al., 2006; Bengio et al., 2007) and Stacked Autoencoders, respectively. Currently, fully supervised, Convolutional Neural Networks (CNN) dominate the leader boards (Krizhevsky et al., 2012; Zeiler and Fergus, 2014; Simonyan and Zisserman, 2014; Ioffe and Szegedy, 2015; He et al., 2015). Their performance increase with respect to the previous decades can largely be attributed to more efficient training methods, advances in hardware such as the employment of many core computing (Cireşan et al., 2011) and most importantly, sheer amounts of annotated training data (Russakovsky et al., 2014).

To the best of our knowledge, Sahiner et al. (1996) were the first to attempt a CNN setup for mammography. Instead of raw images, texture maps were fed to a simple network with two hidden layers, producing two and three feature images respectively. The method gave acceptable, but not spectacular results. Many things have changed since this publication, however, not only with regard to statistical learning, but also in the context of acquisition techniques. Screen Film Mammography (SFM) has made way for Digital Mammography (DM), enabling higher quality, raw images in which pixel values have a well-defined physical meaning and easier spread of large amounts of training data. Given the advances in learning and data, we feel a revisit of CNNs for mammography is more than worthy of exploration.

Work on CAD for mammography (Elter and Horsch, 2009; Nishikawa, 2007; Astley and Gilbert, 2004) has been done since the early nineties but unfortunately, progress has mostly stagnated in the past decade. Methods are being developed on small data sets (Mudigonda et al., 2000; Zheng et al., 2010) which are not always shared and algorithms are difficult to compare (Elter and Horsch, 2009). Breast cancer has two main manifestations in mammography, firstly the presence of malignant soft tissue or masses and secondly the presence of microcalcifications (Cheng and Huang, 2003) and separate systems are being developed for each. Microcalcifications are often small and can easily be missed by oversight. Some studies suggest CAD for microcalcifications is highly effective in reducing oversight (Malich et al., 2006) with acceptable numbers of false positives. However, the merit of CAD for masses is less clear, with research suggesting human errors do not stem from oversight but rather misinterpretation (Malich et al., 2006). Some studies show no increase in sensitivity or specificity with CAD (Taylor et al., 2005) for masses or even a decreased specificity without an improvement in detection rate or characterization of invasive cancers (Fenton et al., 2011; Lehman et al., 2015). We therefore feel motivated to improve upon the state-of-the art.

In previous work in our group (Hupse et al., 2013) we showed that a sophisticated CAD system taking into account not only local information, but also context, symmetry and the relation between the two views of the same breast can operate at the performance of a resident radiologist and of a certified radiologist at high specificity. In a different study (Karssemeijer et al., 2004) it was shown that when combining the judgment of up to twelve radiologists, reading performance improved, providing a lower bound on the maximum amount of information in the medium and suggesting ample room for improvement of the current system.

In this paper, we provide a head-to-head comparison between a CNN and a CAD system relying on an exhaustive set of manually designed features and show the CNN outperforms a state-of-the-art mammography CAD system, trained on a large dataset of around 45,000 images. We will focus on the detection of solid, malignant lesions including architectural distortions, treating benign abnormalities such as cysts or fibroadenomae as false positives. The goal of this paper is *not* to give an optimally concise set of features, but to use a complete set where all descriptors commonly applied in mammography are represented and provide a fair comparison with the deep learning method. As mentioned by Szegedy et al. (2014), success in the past two years in the context of object recognition can in part be attributed to judiciously combining CNNs with classical computational vision techniques. In this spirit, we employ a candidate detector to obtain a set of suspicious locations, which are subjected to further scrutiny, either by the classical system or the CNN. We subsequently investigate to what extent the CNN is still complementary to traditional descriptors by combining the learned representation with features such as location, contrast and patient information, part of which are not explicitly represented in the patch fed to the network. Lastly, a reader study is performed, where we compare the scores of the CNN to experienced radiologists on a patch level.

The rest of this paper is organized as follows. In the next section, we will give details regarding the candidate detection system, shared by both methods. In Section 3, the CNN will be introduced followed by a description of the reference system in Section 4. In Section 5, we will describe the experiments performed and present results, followed by a discussion in Section 6 and conclusion in Section 7.

2. Candidate detection

Before gathering evidence, every pixel is a possible center of a lesion. This approach yields few positives and an overwhelming amount of predominantly obvious negatives. The actual difficult examples could be assumed to be outliers and generalized away, hindering training. Sliding window methods, previously popular in image analysis are recently losing ground in favor of candidate detection (Hosang et al., 2015) such as selective search (Uijlings et al., 2013) to reduce the search space (Girshick et al., 2014; Szegedy et al., 2014). We therefore follow a two-stage classification procedure where in the first stage, candidates are detected and subjected to further scrutiny in a second stage, similar to the pipeline described in Hupse et al. (2013). Rather than class agnostic and potentially less accurate candidate detection methods, we use an algorithm designed for mammographic lesions (Karssemeijer and te Brake, 1996). It operates by extracting 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.

To generate the pixel based training set, we extracted positive samples from a disk of constant size inside each annotated malignant lesion in the training set, to sample the same amount from every lesion size and prevent bias for larger areas. To obtain normal pixels for training, we randomly sampled 1 in 300 pixels from normal tissue in normal images, resulting in approximately 130 negative samples per normal image. The resulting samples were used to train a random forest (Breiman, 2001) (RF) classifier. RFs can be parallelized easily and are therefore fast to train, are less susceptible to overfitting and easily adjustable for class-imbalance and therefore suitable for this task.

To obtain lesion candidates, the RF is applied to all pixel locations in each image, both in the train and test set, generating a likelihood image, where each pixel indicates the estimated suspi-

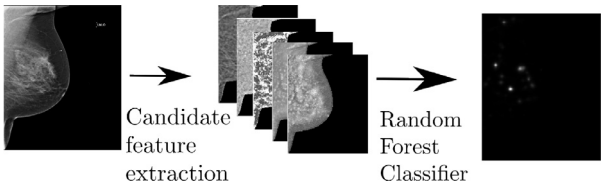
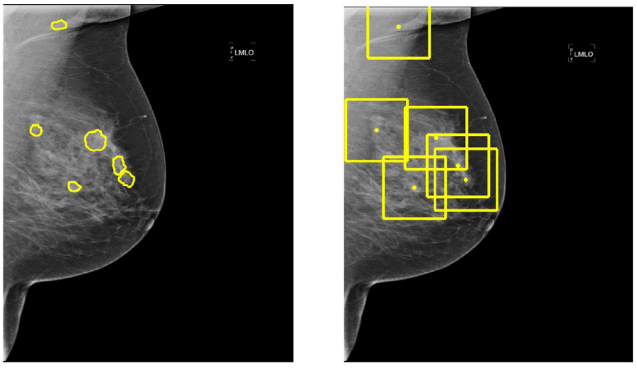


Fig. 1. Illustration of the candidate detection pipeline. A candidate detector is trained using five pixel features and applied to all pixesl in all images, generating a likelihood image. Local optima in the likelihood image are used as seed points for both the reference system and the CNN. (See Fig. 2).



(a) Illustration of segmentations for the reference system.

(b) Illustration of extracted patches for the CNN.

Fig. 2. Two systems are compared. A candidate detector (see Fig. 1) generates a set of candidate locations. A traditional CAD system (left) uses these locations as seed points for a segmentation algorithm. The segmentations are used to compute region based features. The second system based on a CNN (right) uses the same locations as the center of a region of interest.

ciousness. Non-maximum suppression was performed on this image and all optima in the likelihood image are treated as candidates and fed as input to both the reference feature system and the CNN. For the reference system, the local optima in the likelihood image are used as seed points for a segmentation algorithm. For the CNN, a patch centered around the location is extracted. An overview of the first stage pipeline is provided in Fig. 1. Fig. 2 illustrates the generated candidates for both systems.

3. Deep convolutional neural network

In part inspired by human visual processing faculties, CNNs learn hierarchies of filter kernels, in each layer creating a more abstract representation of the data. The term *deep* generally refers to the nesting of non-linear functions (Bengio, 2009). Multi Layered Perceptrons (MLPs) have been shown to be universal function approximators, under some very mild assumptions, and therefore, there is no theoretical limit that prevents them from learning the same mapping as a deep architecture would. Training, however, has been shown, mostly empirically, to be far more efficient in a deep setting and the same function can be represented with fewer parameters. Deep CNN's are currently the most proficient for vision and in spite of the simple mathematics, have been shown to be extremely powerful.

Contemporary architectures roughly comprise *convolutional, pooling* and *fully connected* layers. Every convolution results in a feature map, which is downsampled in the pooling layer. The most common form of pooling is max-pooling, in which the maximum of a neighborhood in the feature map is taken. Pooling induces some translation invariance and downscales the image to reduce the amount of weights with each layer. It also reduces location precision, however, rendering it less suitable for segmentation tasks. The exact merit of fully connected layers is still an open research question, but many studies report an increase in performance with these in the architectures.

If we let $\mathbf{Y}_L^k$ denote the kth feature map of layer L, generated by convolution with kernel $\mathbf{W}^k$, it is computed according to:

$$\mathbf{Y}_L^k = f(\mathbf{W}_L^k * \mathbf{Y}_{L-1} + b_L^k) \tag{1}$$

with $*$ the convolution operator and $f(\cdot)$ a non-linear activation function and b_L^k a bias term. Traditional MLPs use sigmoidal functions to facilitate learning of non-linearly separable problems. However, Rectified Linear Units (ReLU):

$$f(a) = \max(0, a) \tag{2}$$

of activation a, have been shown to be easier to train, since the activation is not squashed by asymptote in the logistic functions (Nair and Hinton, 2010). The parameters $\mathbf{\Theta}$ are typically fit to the data using maximum likelihood estimation:

$$\arg\max_{\mathbf{\Theta}} \mathcal{L}(\mathbf{\Theta}, \mathcal{D}) = \arg\max_{\mathbf{\Theta}} \prod_{n=1}^{N} h(\mathbf{X}|\mathbf{\Theta}) \tag{3}$$

where $h(\mathbf{X}|\mathbf{\Theta})$ produces the posterior probability of sample $\mathbf{X}$. Taking the logarithm and negating it to put it into a minimization framework for convenience, will yield the cross-entropy loss:

$$-\ln[P(\mathcal{D}|\mathbf{\Theta})] = -\sum_{n=1}^{N} yh(\mathbf{X};\mathbf{\Theta}) + (1-y)(1-h(\mathbf{X};\mathbf{\Theta}))] \tag{4}$$

where y indicates the class label. This can be optimized using gradient descent. For large datasets that do not fit in memory and data with many redundant samples, minibatch Stochastic Gradient Descent (SGD) is typically used. Rather than computing the gradient on the entire set, it is computed in small batches. Standard back propagation is subsequently used to adjust weights in all layers.

Although powerful, contemporary architectures are not fully invariant to geometric transformations, such as rotation and scale. Data augmentation is typically performed to account for this.

3.1. Data augmentation

Data augmentation is a technique often used in the context of deep learning and refers to the process of generating new samples from data we already have, hoping to ameliorate data scarcity and prevent overfitting. In object recognition tasks in natural images, simple horizontal flipping is usually only performed, but for tasks such as optical character recognition it has been shown that elastic deformations can greatly improve performance (Simard et al., 2003). The main sources of variation in mammography at a lesion level are rotation, scale, translation and the amount of occluding tissue.

We augmented all positive examples with scale and translation transformations. Full scale or translation invariance is not desired nor required since the candidate detector is expected to find a patch centered around the actual focal point of the lesion. The problem is not completely scale-invariant either: large lesions in a later stage of growth are not simply scaled-up versions of recently emerged abnormalities. The key is therefore to perform the right amount of translation and scaling in order to generate realistic lesion candidates. To this end, we translate each patch in the training set containing an annotated malignant lesion 16 times by adding values sampled uniformly from the interval $[-25, 25]$ (0.5 cm) to the lesion center and scale it 16 times by adding values from the interval $[-30, 30]$ (0.6 cm) to the top left and bottom right of the bounding box. After this, all patches including the normals were rotated using simple flipping actions, which can be computed on the fly to generate three more samples. This results

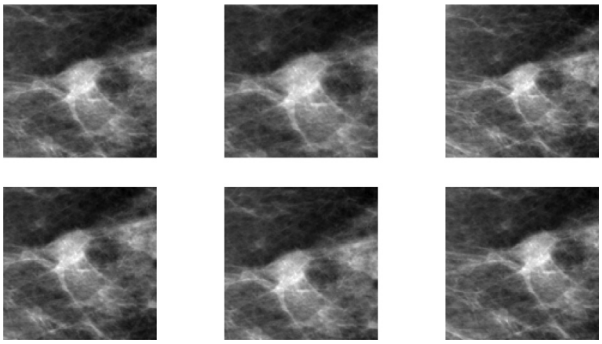


Fig. 3. Examples of scaling and translation of the patches. The top left image is the original patch, the second and third image of the top row examples of the smallest and largest scaling employed. The bottom row indicates the extrema in the range of translation used.

in $(1 + 16 + 16)4 = 132$ patches per positive lesions and 4 per negative. Examples of the range of scaling and translation augmentation are given in Fig. 3.

4. Reference system

The large majority of CAD systems rely on some form of segmentation of the candidates on which region based features are computed. To this end, we employ the mass segmentation method proposed by Timp and Karssemeijer (2004), which was shown to be superior to other methods (region growing (te Brake and Karssemeijer, 2001) and active contour segmentation (Kupinski and Giger, 1998)) on their particular feature set. The image is transformed to a polar domain around the center of the candidate and dynamic programming is used to find an optimal contour, subject to the constraint that the path must start and end in the same column in order to generate a closed contour in the Cartesian domain. A cost function incorporating a deviation from the expected Grey level, edge strength and size terms is used to find an optimal segmentation. One of the problems with this method and many knowledge driven segmentation methods for that matter, is that it is conditioned on a *false prior*: the size constraint is based on data from malignant lesions. When segmenting a candidate, we therefore implicitly assume that this is a malignant region, inadvertently driving the segmentation into a biased result. Many of the manual features described below rely on a precise segmentation but in the end, it is an intermediate problem. For a stand-alone application, we are interested to provide the patient with an accurate diagnosis, not the exact delineation. A huge advantage of CNNs is that no segmentation is required and patches are fed without any intermediate processing.

After segmentation, we extract a set of 74 features. These can broadly be categorized into *pixel level features*, used by the candidate detector, *contrast features*, capturing the relation between the attenuation coefficients inside and outside the region, *texture features* describing relations between pixels within the segmented region, *geometry features* summarizing shape and border information *location* features, indicating where the lesion is with respect to some landmarks in the breast, *context* features, capturing information about the rest of the breast and other candidates and *patient* features, conveying some of the subjects background information.

4.1. Candidate detector features

As a first set of descriptors, we re-use the five features employed by the candidate detector, which has been shown to be beneficial in previous work in our group. On top of this, we compute the mean of the four texture features within the segmented boundary and add the output of the candidate detector at the found optimum. This gives us a set of nine outputs we call candidate detector features.

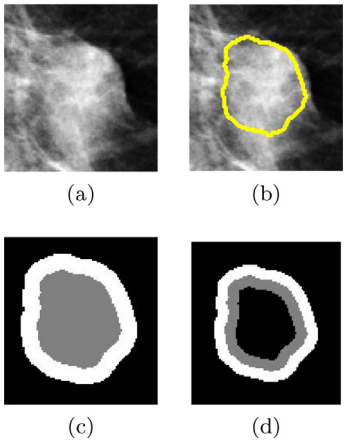


Fig. 4. A lesion (a), its segmentation (b), areas used for computing contrast features (c) and areas used for computing margin contrast (d).

4.2. Contrast features

When talking to a radiologist, a feature that is often mentioned is how well a lesion is separated from the background. Contrast features are designed to capture this. To compute these, we apply a distance transform to the segmented region and compare the inside of the segmentation with a border around it. The distance d to the border of the segmentation is determined according to:

$$d = \rho\sqrt{A\pi} \tag{5}$$

with A the area of the segmented lesion. An illustration is provided in Fig. 4. An important nuisance in this setting is the tissue surrounding the lesion. In previous work, we have derived two model based features, designed to be invariant to this factor (Kooi and Karssemeijer, 2014), which were also normalized for size of the lesion. The *sharpness* of the border of the lesion is also often mentioned by clinicians. To capture this, we add two features: the acutance (Rangayyan et al., 1997) and margin contrast, the different between the inside and outside of the segmentation, using a small margin. Illustrations of contrast features are provided in Fig. 4. Other contrast features described in te Brake et al. (2000) were added to give a set of 12 features.

4.3. Texture features

The presence of holes in the candidate lesion often decrease their suspiciousness, since tumours are solid, with possibly the exception of lobular carcinoma. To detect this, we added the two isodensity features proposed by te Brake et al. (2000). Linear structures within a lesion can indicate an unfortunate projection rather then cancer, for which we used four linear texture features as described by the same authors (te Brake et al., 2000). On top of this we added two features based on the second order gradient image of the segmented lesion. The image was convolved with second order Gaussian derivative filters and the optimal location in scale space was selected for each pixel. We subsequently took the first and second moment of the segmented lesion of the maximum magnitude, which is expected to be high for lesions with much line structure. Secondly, we computed gradient coocurence, by counting the number of times adjacent pixels have the same orientation. Ten less biophysical features in the form of Haralick features

(Haralick et al., 1973) at two different scales (*entropy, contrast, correlation, energy* and *homogeneity*) were added to give a set of 21 texture descriptors.

4.4. Geometrical features

Regularity of the border of a lesion is often used to classify lesions. Again, expedient computation relies heavily on proper segmentations. Nevertheless, we have incorporated five simple topology descriptors as proposed by Peura and Iivarinen (1997) in the system. These are *eccentricity, convexity, compactness, circular variance* and *elliptic variance*. In order to capture more of the 3D shape, we extended these descriptors to also work with 3 dimensions. The lesion was smoothed with a Gaussian kernel first and *3D eccentricity*: the ratio between the largest and smallest eigenvalue of the point cloud, *3D compactness*: the ratio of the surface area to the volume, *spherical deviance*, the average deviation of each point from a sphere and *elliptical deviance*: the average deviation of each point to an ellipse fitted to the point cloud were computed. Since convex hull algorithms in 3D suffer from relatively high computational complexity, this was not extended. Op top of this, we added a feature measuring reflective symmetry. The region is divided into radial and angular bins and average difference pixel intensity between opposing bins is summed and normalized by the size of the region. Lastly the area of the segmented region is added, giving us a set of 10 geometric features.

4.5. Location features

Lesions are more likely to occur in certain parts of the breast than others and other structures such as lymph nodes are more common in the pectoralis than in other parts of the breast. To capture this, we use a simple coordinate system. The area of the breast and pectoral muscle are segmented using thresholding and a polynomial fit. We subsequently estimate the nipple location by taking the largest distance to the chest wall and a central landmark in the chest wall is taken as the row location of the center of gravity. From this, we extract: (1) the distance to the nipple (2) the same, but normalized for the size of the breast, (3) the distance to the chest wall and (4) the fraction of the lesion that lies in the pectoral muscle.

4.6. Context features

To add more information about the surroundings of the lesion, we added three context features as described by Hupse and Karssemeijer (2009). The features again make use of the candidate detector and assume the posterior of pixels in the rest of the breast convey some information about the nature of the lesion in question. The first feature averages the output around the lesion, the second in a band at a fixed distance from the nipple and a third takes the whole segmented breast into account. On top of this, we added the posterior of the candidate detector, normalized by the sum of the top three and top five lesions in the breast, to give us five context features in total.

4.7. Patient features

Lastly, we added the age of the patient, which is an important risk factor. From the age, we also estimate the screening round by subtracting 50 (the age at which screening starts in The Netherlands) and dividing by 2 (the step size of the screening). This gives us two features.

Note that the last three sets of features provide information outside of the patch fed to the CNN. Even if the network is able to exploit all information in the training set, these could still supply complementary information regarding the nature of the lesion.

Table 1
Overview of the data. Pos refers to the amount of malignant lesions and neg to the amount of normals.

	Cases		Exams		Images		Candidates	
	Pos	Neg	Pos	Neg	Pos	Neg	Pos	Neg
Train	296	6433	358	11,780	634	39,872	634	213,450
Valid.	35	710	42	1247	85	4218	85	19,460
Test	124	2064	124	5317	271	18,182	271	180,777

5. Experiments

5.1. Data

The mammograms used were collected from a large scale screening program in The Netherlands (*bevolkingsonderzoek midden-west*) and recorded using a Hologic Selenia digital mammography system. All tumours are biopsy proven malignancies and annotated by an experienced reader. Before presentation to a radiologist, the manufacturer applies some processing to optimize it for viewing by a human. To prevent information loss and bias, we used the raw images instead and only applied a log transform which results in pixel values being linearly related to the attenuation coefficient. Images were scaled from 70 micron to 200 for faster processing. Structure important for detecting lesions occur at larger scales and therefore this does not cause any loss of information.

An overview of the data is provided in Table 1. With the term case, we refer to all screening images recorded from a single patient. Each case consists of several exams taken at typically a two year interval and each exam typically comprises four views, two of each breast, although these numbers vary: some patients skip a screening and for some exams only one view of each breast is recorded. For training and testing, we selected all regions found by the candidate detector. The train, validation and test set were all split on a patient level to prevent any bias. The train and validation set comprise 44,090 mammographic views, from which we used 39,872 for training and 4218 for validation. The test set consisted of 18,182 images of 2064 patients with 271 malignant annotated lesions. A total of 30 views from 20 exams in the test set contained an interval cancer that was visible in the mammogram or were taken prior to a screen detected cancer, with the abnormality already visible.

Before patch extraction in the CNN system, we segmented all lesions in the training set in order to get the largest possible lesion and choose the patch size with an extra margin resulting in patches of size 250 $\times$ 250 (5 $\times$ 5 cm). The pixel values in the patches were scaled using simple min-max scaling, with values calculated over the whole training set. We experimented with scaling the patches locally, but this seemed to perform slightly though not significantly worse on the validation set. All interpolation processes were done with bilinear interpolation. Since some candidates occur at the border of the imaged breast, we pad the image with zeros. Negative examples were only taken from normal images. Annotated benign samples such as cysts and fibroadenomae were removed from the training set. However, not all benign lesions in our data are annotated and therefore some may have ended in the train or validation set as negatives. After augmentation, the train set consisted of $334,752$ positive patches and $853,800$ negatives. When combining the train and validation set, this amounts to $379,632$ positive and $931,640$ negative patches.

5.2. Training and classification details

For the second stage classification, we have experimented with several classifiers (SVMs with several different kernels, Gradient

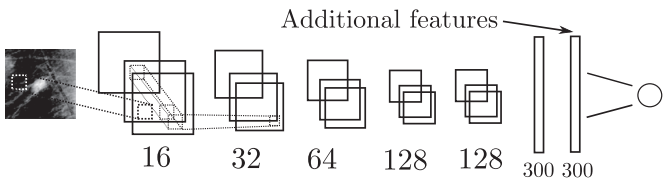


Fig. 5. Illustration of the network architecture, The numbers indicate the amount of kernels used. We employ a scaled down version of the VGG model. To see the extent to which conventional features can still help, the network is trained fully supervised and the learned features are subsequently extracted from the final layer and concatenated with the manual features and retrained using a second classifier.

Boosted Trees, MLPs) on a validation set, but found in nearly all circumstances the random forest performed similar or better than others. To counteract the effect of class imbalance, trees in the RF were grown using the balance corrected Gini criterion for splitting and in all situations we used 2000 estimators and the square root heuristic for the maximum number of features. The maximum depth was cross-validated using 8 folds. We employed class weights inversely proportional to the distribution in the particular bootstrap sample. The posterior probability output by the RF was calculated as a mean of the estimated classes. The systems are trained using at most the ten most suspicious lesions per image found by the candidate detector, during testing no such threshold is applied to obtain highest sensitivity.

We implemented the network in Theano (Bergstra et al., 2010) and pointers provided by Bengio (2012) were followed and very helpful. We used OxfordNet-like architectures (Simonyan and Zisserman, 2014) with 6 convolutional layers of {16, 32, 64, 128, 128} with 3×3 kernels and 2×2 max-pooling on all but the fourth convolutional layer. A stride of 1 was used in all convolutions. Two fully connected layers of 300 each were added. An illustration of the network is provided in Fig. 5.

We employed Stochastic Gradient Descent (SGD) with RMSProp (Dauphin et al., 2015), an adaption of R-Prop for SGD with Nesterov momentum (Sutskever et al., 2013). Drop-out (Srivastava et al., 2014) was used on the fully connected layers with $p = 0.5$. We used the MSRA (He et al., 2015) weight filler, a learning rate of 5×10^{-5} with a weight decay of 5×10^{-5}. To battle the strong class imbalance, positive samples were presented multiple times during an epoch, keeping a 50/50 positive/negative ratio in each minibatch. Alternatively, the loss function could be weighted, but we found this to perform worse, we suspect this is because rebalancing maintains a certain diversity in the minibatch. All hyperparameters were optimized on a validation set and the CNN was subsequently retrained on the full training set using the found parameters. All test patches were also augmented using the same augmentation scheme. On the validation set, this gave a small improvement. The best validation AUC was 0.90.

5.3. ROC analysis

To first get an understanding of how well each feature set performs individually, we trained different RFs for each feature set and applied them separately to the test set. In all cases, the training procedure as described above was used. AUC values along with a 95% confidence interval, acquired using bootstrapping (Efron, 1979; Bornefalk and Hermansson, 2005) with 5000 bootstrap samples are shown in Table 2.

The CNN was compared to the reference system with equal amount of information (i.e., excluding location, context and patient information) to get a fair performance comparison. Fig. 6 shows a plot of the mean curves along with the 95% confidence interval obtained after bootstrapping. Results were not found to be significantly different $p = 0.2$ on the full ROC. Fig. 7 shows a plot comparing the CNN with data augmentation to the network without data augmentation and with data augmentation and added manual features. Again bootstrapping was used to obtain significance. It is clear that the proposed data augmentation methods contributes greatly to the performance, which was also found to be significant ($p \ll 0.05$).

To combine the CNN with other descriptors, we extracted the features from the last fully connected layer and appended the other set (see Fig. 5). For each augmented patch, the additional features were simply duplicated. Table 3 shows results of the CNN

Table 2
Overview of results of individual feature sets along the 95% confidence interval (CI) obtained using 5000 bootstraps.

Feature group	AUC	CI
Candidate detector	0.858	[0.827, 0.887]
Contrast	0.787	[0.752, 0.817]
Texture	0.718	[0.681, 0.753]
Geometry	0.753	[0.721, 0.784]
Location	0.686	[0.651, 0.719]
Context	0.816	[0.781, 0.850]
Patient	0.651	[0.612, 0.688]
Equal information	0.892	[0.864, 0.918]
All	0.906	[0.881, 0.929]

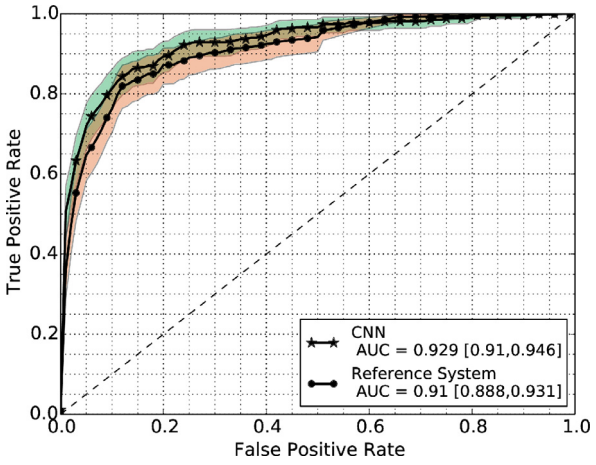


Fig. 6. Comparison of the CNN with the reference system using equal information, i.e., only information represented in the patch used by the CNN, excluding context, location and patient information.

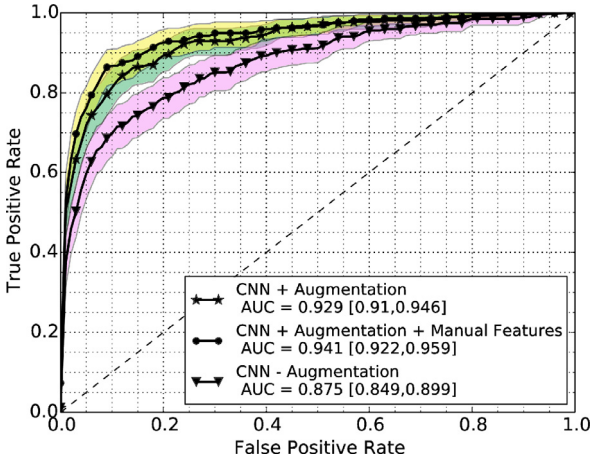


Fig. 7. Comparison of the CNN without any augmentation, with augmentation and with added manual features.

Table 3
Overview of results of the CNN combined with individual feature sets.

Feature group added to CNN	AUC	CI
CNN Only	0.929	[0.897, 0.938]
Candidate detector	0.938	[0.919, 0.955]
Contrast	0.931	[0.91, 0.949]
Texture	0.933	[0.912, 0.950]
Geometry	0.928	[0.907, 0.946]
Location	0.933	[0.913, 0.950]
Context	0.934	[0.914, 0.952]
Patient	0.929	[0.908, 0.947]
All	0.941	[0.922, 0.958]

Table 4
AUC values obtained when training the network on subsets of malignant lesions in the training set, keeping the same amount of normals.

Data Augmentation	60%	All
With	0.842	0.929
Without	0.685	0.875

combined with different feature sets, again with confidence interval acquired by bootstrapping with 5000 samples.

To investigate the degree to which a large data set is really needed, we trained several networks on subsets, removing 40% of the malignant lesions. Results are provided in Table 4. Since the differences are rather large, we did not perform significance testing. For all settings, we optimized the learning rate but kept all other hyperparameters equal to the ones found to be optimal for the full training set.

5.4. FROC analysis

In practice, a CAD system should ideally be operating at a referral rate similar to that of a radiologists. To get a better understanding of the system's performance around this operating point, we compute the Partial Area Under the Curve (PAUC) on a log scale:

$$PAUC = \frac{1}{\ln[1] - \ln[0.01]} \int_{0.01}^{1} \frac{s(f)}{f} \mathrm{d}f \tag{6}$$

and generate Free Receiver Operator Characteristic (FROC) curves, to illustrate the numbers of false positives per image.

Plots of the FROCs of the full reference system (last line in Table 2), the CNN only and the CNN plus manual features are shown in Figs. 8 and 9. To further investigate which features are helpful at high specificity, we compute PAUC for each feature set individually. Results are shown in Table 5. We see a significant difference comparing the CNN with additional features to the reference system $P = 0.015$ on a lesion level and $P = 0.0002$ on a case level.

5.5. Human performance

In previous work in our group, performance of the CAD system was compared to the performance of a radiologists at an exam level, a collection of four images, which contains more information than only a patch, such as context in the mammogram, symmetrical difference between two breast, the relation between the CC and MLO views. To get a better understanding of how close the CNN is to human performance on a patch level and how much more room there is for improvement in this sub part of the pipeline, we performed a study where we measured the performance of experienced readers on a patch level, providing the reader with the same

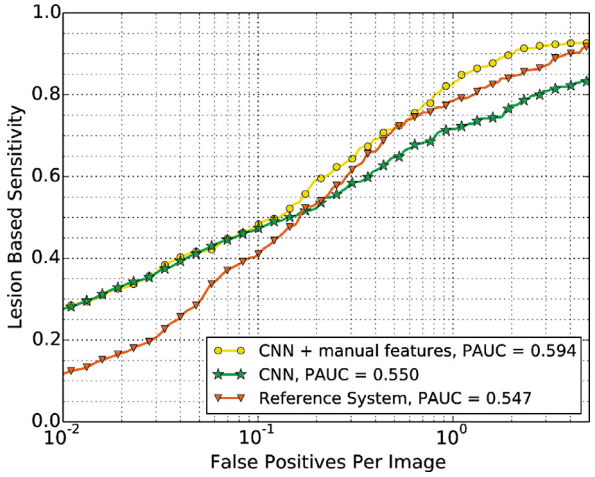


Fig. 8. Lesion based FROC of the three systems. Please note that this concerns the full reference system, where context, location and patient features are incorporated.

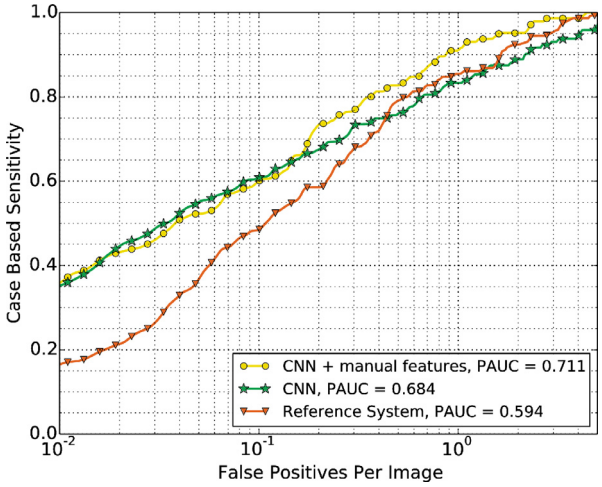


Fig. 9. Case based FROC of the three systems. In areas of high specificity, the CNN and the addition of manual features is particularly useful. Please note that this concerns the full reference system, where context, location and patient features are incorporated.

information as the CNN. The group of readers consisted of one experienced reader (non-radiologist) and two experienced certified radiologists. To get an idea of the performance that can at least be obtained on this set, the mean of the three readers was also computed by simply averaging the scores that each of the three readers assigned to each patch.

Patches were extracted from the mammogram processed by the manufacturer for optimal viewing and were shown at a normal computer screen at a resolution of 200 micron. Microcalcifications are difficult to see in this setting, but all structures relevant for soft tissue lesions are intact and readers did not report difficulties. The readers were provided with a slider and instructed to score the patch between zero and one hundred based on their assessment of the suspiciousness of the patch.

As a test set, we used all masses that were used in Hupse et al. (2013) and selected an equal amount of negatives, that were considered the most difficult by the candidate detector, resulting in 398 patches. This gives a representative set of difficult samples and allows for larger differences between readers and the CNN, but is biased towards a set difficult for the reference system, which was therefore left out of the comparison (obtained AUC was 0.64 on this set). Fig. 12 shows the ROC curves resulting from the

Table 5
Partial Area under the FROC of different systems. P-values are referring to the comparison between the CNN with additional features and the CNN without the specific feature group. In this case, the reference system is the *full* system, including context, location and patient information.

	Lesion	Case	P, lesion	P, case
CNN	0.550	0.684	1	1
Reference system	0.547	0.594	0.451	0.013
CNN + candidate det.	0.590	0.701	< **0.0001**	**0.026**
CNN + contrast	0.571	0.704	**0.011**	0.0758
CNN + texture	0.574	0.705	**0.0062**	0.067
CNN + topology	0.561	0.700	**0.0286**	0.132
CNN + location	0.576	0.707	**0.0038**	0.0516
CNN + context	0.578	0.700	**0.0028**	0.121
CNN + patient	0.576	0.704	**0.0034**	0.0784
CNN + all features	0.594	0.711	< **0.001**	**0.04**

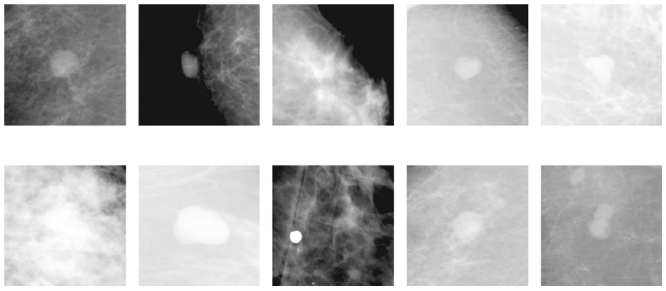


Fig. 10. Top misclassified negatives by the CNN. The second sample in the first row is simply the nipple and the third sample in the second row displays fat necrosis. Both are obviously normal patches and are filtered out using additional feature sets.

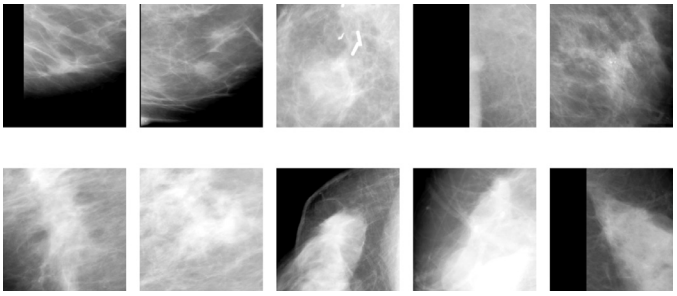


Fig. 11. Top misclassified positives by the CNN, most samples are very large lesion unlikely to be found in the screening population and therefore under represented in the training set.

reader study. Again, to test significance we used bootstrapping and two sided testing to get a significance score. We found no significant difference between the CNN and any of the readers: CNN vs reader 1: $p = 0.1008$, CNN vs reader 2: $p = 0.6136$, CNN vs reader 3: $p = 0.64$, but found a significant difference between the CNN and the mean of the human readers ($p = 0.001$).

6. Discussion

To get more insight into the performance of the network, examples of the top misclassified positives and negatives are shown in Fig. 11 and 10 respectively. A large part of the patches determined as suspicious by the network are benign abnormalities such as cysts and fibroadenomae or normal structures such as lymph nodes or fat necrosis. Cysts and lymph nodes can look relatively similar to masses. These strong false positives occur due to the absence of benign lesions in our training set. In the future we plan

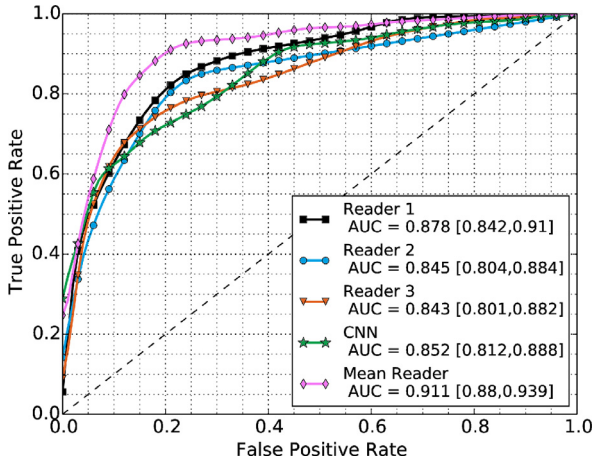


Fig. 12. Comparison between the CNN and three experienced readers on a patch level.

to add these to the training set and perform three-class classification or train a separate network to discriminate these lesions properly.

The majority of 'misclassified' positives are lesions ill-represented in the training data, either very subtle or extremely large. When using CAD as a second reader, these will not influence the referral decision much, as they are clearly visible to a human, but when using the computer as an independent reader, these issues need to be solved. In preliminary experiments, we have seen that many of these misclassifications can be prevented by considering the contralateral breast and plan to work on this in the future.

From the results in Tables 3 and 2 we can see that individually, apart from the candidate detector, contrast and context are useful features. Although age and screening round are some of the most important risk factors, we do not see clear improvements when added as features, which is slightly disappointing. To get training data, we took negative patches only from normal images, but not only from normal exams, to get as many data points as possible. A possible explanation for the disappointing performance may be that the relation between age and cancer is more difficult to learn in the setting, since it is a relation that exist on an exam level.

To add features, we have used a second classification stage. This has the advantage it is easy to evaluate which features add information, without retraining a network and re-optimizing the parameters, which can take several weeks to do properly. On top of this, the learned feature representation of the CNN is the same in all situations, rendering comparison more reliable. A major disadvantage, however, is that the training procedure is rather complicated. Other more elegant methods such as coding features as a second channel, as done by Maddison et al. (2014) or adding the features in one of the fully connected layers of the network during training could be better strategies and we plan to explore this in future work.

We have made use of a more shallow and scaled down version of the networks proposed by Simonyan and Zisserman (2014), who obtain best performance on ImageNet with a 19 layer architecture with four times the amount of kernels in each layer. In initial experiments, we have worked with Alexnet-like architectures, which performed worse on our problem, obtaining and AUC of around 0.85 on the validation set. We have also experimented with deeper networks and increasing the amount of kernels, but found no significant improvement on the validation set (0.896 vs 0.897 of the network with larger capacity and 0.90 of 9 layer network). We suspect that with more data, larger capacity

networks can become beneficial. The problem could be less complex than classifying natural images since it concerns a two-class classification in the current setting and we are dealing with gray scale images, contrary to the thousands of classes and RGB data in ImageNet (Russakovsky et al., 2014). Therefore, more shallow and lower capacity networks than the one found optimal for natural images could suffice for this particular problem.

In our work, we made extensive use of data augmentation in the form of simply geometric transformations. We have also experimented with full rotation, but this creates lesions not expected during testing, due to the zero padding. This could be prevented using test time augmentation, but when used in a sliding window fashion this is not convenient. The ROC curves in Fig. 7 show a clear increase in performance for the full data set. The results in Table 4 show the current data augmentation scheme improves performance for large amounts of data but not for small amounts of data. We suspect in the latter setting, the network overfits and more regularization is needed. These results may be different when fully optimizing the architecture and augmentation procedure for each setting individually. More research is needed to draw clear conclusions. However effective, data augmentation is a rather computationally costly procedure. A more elegant approach would be to add the invariance properties in the network architecture, which is currently being investigated in several papers (Gens and Domingos, 2014; Jaderberg et al., 2015). On top of the geometric transforms, occluding tissue is an important source of variance, which is more challenging to explicitly code in the network architecture. In future work, we plan to explore simulation methods for this.

In this work, we have employed a previously developed candidate detector. This has two main advantages: (1) it is fast and accurate (2) the comparison with the traditional CAD system is straightforward and fair, since exactly the same candidate locations are trained with and evaluated on. The main disadvantage is that the sensitivity is not hundred percent, which causes lesions to be missed, although the case-based performance is close to optimal. In future work, we plan to explore other methods, such as the strategy put forth by Cireşan et al. (2013), to train the system end-to-end. This will make training and classification less cumbersome and has the potential to increase the sensitivity of the system.

In this work we have compared the CNN to a state-of-the art CAD system (Hupse et al., 2013), which was combined with several other features commonly used in the mammography CAD literature. A random forest was subsequently used, that performs feature selection during its training stage. We think the feature set we used is sufficiently exhaustive to include most features commonly used in literature and therefore think similar conclusions hold for other state-of-the art CAD systems. To the best of our knowledge, the Digital Database of Screening Mammography (DDSM) is the only publicly available data set, which comprises of digitized screen film mammograms. Since almost all screening centers have migrated to digital mammography, we have elected not to run our system on this data set, because we think the clinical relevance is arguable. On top of this, since this entails a *transfer learning* problem, the system may require retraining to adapt to the older modality.

The reader study illustrates the network is not far from the radiologists performance, but still substantially below the mean of the readers, suggesting a large performance increase is still possible. We suspect that some other augmentation methods as discussed above could push the network a bit further, but expect more training data, when it becomes available will be the most important factor. Also, we feel still employing some handcrafted features that specifically target weaknesses of the CNN may be a good strategy and may be more pragmatic and effective than adding thousands of extra samples to the training set.

7. Conclusion

In this paper we have shown that a deep learning model in the form of a Convolutional Neural Network (CNN) trained on a large data set of mammographic lesions outperforms a state-of-the art system in Computer Aided Detection (CAD) and therefore has great potential to advance the field of research. A major advantage is that the CNN learns from data and does not rely on domain experts, making development easier and faster. We have shown that the addition of location information and context can easily be added to the network and that several manually designed features can give some small improvements, mostly in the form of 'common sense': obviously false negatives will no longer be considered as such. On top of this, we have compared the CNN to a group of three experienced readers on a patch level, two of which were certified radiologist and have show that the human readers and CNN have similar performance.

Acknowledgements

This research was funded by grant KUN 2012-557 of the Dutch Cancer Society and supported by the Foundation of Population Screening Mid West.

References

Astley, S.M., Gilbert, F.J., 2004. Computer-aided detection in mammography. Clinic. Radiol. 59, 390–399.
Bengio, Y., 2009. Learning deep architectures for ai. Found.Trends Mach. Learn. 2, 1–127.
Bengio, Y., 2012. Practical recommendations for gradient-based training of deep architectures. In: Neural Networks: Tricks of the Trade. Springer, pp. 437–478.
Bengio, Y., Courville, A., Vincent, P., 2013. Representation learning: A review and new perspectives. IEEE Trans. Pattern Anal. Mach. Intell. 35, 1798–1828.
Bengio, Y., Lamblin, P., Popovici, D., Larochelle, H., 2007. Greedy layer-wise training of deep networks. In: Advances in Neural Information Processing Systems, pp. 153–160.
Bergstra, J., Breuleux, O., Bastien, F., Lamblin, P., Pascanu, R., Desjardins, G., Turian, J., Warde-Farley, D., Bengio, Y., 2010. Theano: a CPU and GPU math expression compiler. In: Proceedings of the Python for Scientific Computing Conference (SciPy), p. 3.
Bornefalk, H., Hermansson, A.B., 2005. On the comparison of froc curves in mammography CAD systems. Med. Phys. 32, 412–417.
te Brake, G.M., Karssemeijer, N., 2001. Segmentation of suspicious densities in digital mammograms. Med. Phys. 28, 259–266.
te Brake, G.M., Karssemeijer, N., Hendriks, J.H., 2000. An automatic method to discriminate malignant masses from normal tissue in digital mammograms. Phys. Med. Biol. 45, 2843–2857.
Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.
Broeders, M., Moss, S., Nyström, L., Njor, S., Jonsson, H., Paap, E., Massat, N., Duffy, S., Lynge, E., Paci, E., 2012. The impact of mammographic screening on breast cancer mortality in europe: a review of observational studies. J. Med. Screening 19, 1425.
Cheng, S.C., Huang, Y.M., 2003. A novel approach to diagnose diabetes based on the fractal characteristics of retinal images. IEEE Trans. Inf. Technol. Biomed. 7, 163–170.
Cireşan, D.C., Giusti, A., Gambardella, L.M., Schmidhuber, J., 2013. Mitosis detection in breast cancer histology images with deep neural networks. In: Medical Image Computing and Computer-Assisted Intervention, pp. 411–418.
Cireşan, D.C., Meier, U., Masci, J., Maria Gambardella, L., Schmidhuber, J., 2011. Flexible, high performance convolutional neural networks for image classification. In: International Joint Conference on Artificial Intelligence, p. 1237.
Cireşan, D.C., Meier, U., Masci, J., Schmidhuber, J., 2012. Multi-column deep neural network for traffic sign classification. Neural Netw. 32, 333–338.
Dauphin, Y. N., de Vries, H., Chung, J., Bengio, Y., 2015. Rmsprop and equilibrated adaptive learning rates for non-convex optimization. arXiv:150204390.
Doi, K., 2005. Current status and future potential of computer-aided diagnosis in medical imaging. British J. Radiol. 78 Spec No 1, S3–S19.
Doi, K., 2007. Computer-aided diagnosis in medical imaging: historical review, current status and future potential. Comput. Med. Imag. Graph. 31, 198–211. PMID: 17349778.
Efron, B., 1979. Bootstrap methods: Another look at the jackknife. Annals Stat. 7, 1–26.
Elter, M., Horsch, A., 2009. Cadx of mammographic masses and clustered microcalcifications: a review. Med. Phys. 36, 2052–2068.
Fenton, J.J., Abraham, L., Taplin, S.H., Geller, B.M., Carney, P.A., D'Orsi, C., Elmore, J.G., Barlow, W.E., 2011. Effectiveness of computer-aided detection in community mammography practice. J. Nation. Cancer Inst. 103, 1152–1161.

Fukushima, K., 1980. Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biol. Cybern. 36, 193–202.
Gens, R., Domingos, P. M., 2014. Deep symmetry networks. 2537–2545.
Giger, M.L., Karssemeijer, N., Armato, S.G., 2001. Computer-aided diagnosis in medical imaging. IEEE Trans. Med. Imag. 20, 1205–1208.
van Ginneken, B., Schaefer-Prokop, C., Prokop, M., 2011. Computer-aided diagnosis: how to move from the laboratory to the clinic. Radiology 261, 719–732.
Girshick, R., Donahue, J., Darrell, T., Malik, J., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In: Computer Vision and Pattern Recognition. IEEE, pp. 580–587.
Haralick, R.M., Shanmugam, K., Dinstein, I., 1973. Textural features for image classification. IEEE Trans. Syst. Man Cybern. 3, 610–621.
He, K., Zhang, X., Ren, S., Sun, J., 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. Comput. Vis. Pattern Recognit. 1026–1034. 1502.01852v1.
Hinton, G.E., Osindero, S., Teh, Y.W., 2006. A fast learning algorithm for deep belief nets. Neural Comput. 18, 1527–1554.
Hosang, J., Benenson, R., Dollár, P., Schiele, B., 2015. What makes for effective detection proposals? arXiv:150205082.
Hupse, R., Karssemeijer, N., 2009. Use of normal tissue context in computer-aided detection of masses in mammograms. IEEE Trans. Med. Imag. 28, 2033–2041.
Hupse, R., Samulski, M., Lobbes, M., den Heeten, A., Imhof-Tas, M.W., Beijerinck, D., Pijnappel, R., Boetes, C., Karssemeijer, N., 2013. Standalone computer-aided detection compared to radiologists' performance for the detection of mammographic masses. Eur. Radiol. 23, 93–100.
Ioffe, S., Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv:150203167.
Jaderberg, M., Simonyan, K., Zisserman, A., Kavukcuoglu, K., 2015. Spatial transformer networks. arXiv preprint arXiv:150602025.
Karssemeijer, N., te Brake, G., 1996. Detection of stellate distortions in mammograms. IEEE Trans. Med. Imag. 15, 611–619.
Karssemeijer, N., Otten, J.D., Roelofs, A.A.J., van Woudenberg, S., Hendriks, J.H.C.L., 2004. Effect of independent multiple reading of mammograms on detection performance. In: Medical Imaging, pp. 82–89.
Kooi, T., Karssemeijer, N., 2014. Invariant features for discriminating cysts from solid lesions in mammography. In: Breast Imaging. Springer, pp. 573–580.
Krizhevsky, A., Sutskever, I., Hinton, G., 2012. Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems 25, pp. 1097–1105.
Kupinski, M.A., Giger, M.L., 1998. Automated seeded lesion segmentation on digital mammograms. IEEE Trans. Med. Imag> 17, 510–517.
LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436–444.
Lecun, Y., Bottou, L., Bengio, Y., Haffner, P., 1998. Gradient-based learning applied to document recognition. Proc. IEEE 86, 2278–2324.
Lehman, C.D., Wellman, R.D., Buist, D.S.M., Kerlikowske, K., Tosteson, A.N.A., Miglioretti, D.L.,B.C.S.C., 2015. Diagnostic accuracy of digital screening mammography with and without computer-aided detection.. JAMA Intern Med 175, 1828–1837.
Maddison, C. J., Huang, A., Sutskever, I., Silver, D., 2014. Move evaluation in go using deep convolutional neural networks. arXiv:14126564.
Malich, A., Fischer, D.R., Böttcher, J., 2006. CAD for mammography: the technique, results, current role and further developments. Eur. Radiol. 16, 1449–1460.
Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M., Fidjeland, A.K., Ostrovski, G., 2015. Human-level control through deep reinforcement learning. Nature 518, 529–533.
Mudigonda, N.R., Rangayyan, R.M., Desautels, J.E., 2000. Gradient and texture analysis for the classification of mammographic masses. IEEE Trans. Med. Imag. 19, 1032–1043.
Nair, V., Hinton, G.E., 2010. Rectified linear units improve restricted boltzmann machines. In: International Conference on Machine Learning, pp. 807–814.
Nishikawa, R.M., 2007. Current status and future directions of computer-aided diagnosis in mammography. Comput. Med. Imag. Graph. 31, 224–235.
Peura, M., Iivarinen, J., 1997. Efficiency of simple shape descriptors. In: Proceedings of the third international workshop on visual form, p. 451.
Rangayyan, R.M., El-Faramawy, N.M., Desautels, J.E.L., Alim, O.A., 1997. Measures of acutance and shape for classification of breast tumors. IEEE Trans. Med. Imag. 16, 799–810.
Rao, V.M., Levin, D.C., Parker, L., Cavanaugh, B., Frangos, A.J., Sunshine, J.H., 2010. How widely is computer-aided detection used in screening and diagnostic mammography? J. Am. College Radiol. 7, 802–805.
Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., 2014. Imagenet large scale visual recognition challenge. Int. J. Comput. Vis. 115, 1–42.
Sahiner, B., Chan, H.P., Petrick, N., Wei, D., Helvie, M.A., Adler, D.D., Goodsitt, M.M., 1996. Classification of mass and normal breast tissue: a convolution neural network classifier with spatial domain and texture images. IEEE Trans. Med. Imag. 15, 598–610.
Schmidhuber, J., 2015. Deep learning in neural networks: an overview. Neural Netw. 61, 85–117.
Simard, P.Y., Steinkraus, D., Platt, J.C., 2003. Best practices for convolutional neural networks applied to visual document analysis. In: Document Analysis and Recognition, pp. 958–963.
Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv:14091556.
Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: A simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. 15, 1929–1958.
Sutskever, I., Martens, J., Dahl, G., Hinton, G., 2013. On the importance of initialization and momentum in deep learning. In: International Conference on Machine Learning, pp. 1139–1147.
Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2014. Going deeper with convolutions. arXiv:14094842v1. 1409.4842v1.
Tabar, L., Yen, M.F., Vitak, B., Chen, H.H.T., Smith, R.A., Duffy, S.W., 2003. Mammography service screening and mortality in breast cancer patients: 20-year follow-up before and after introduction of screening. Lancet 361, 1405–1410.
Taylor, P., Champness, J., Given-Wilson, R., Johnston, K., Potts, H., 2005. Impact of computer-aided detection prompts on the sensitivity and specificity of screening mammography. Health Technol. Assess. 9, iii,1–iii,58.
Timp, S., Karssemeijer, N., 2004. A new 2D segmentation method based on dynamic programming applied to computer aided detection in mammography. Med. Phys. 31, 958–971.
Uijlings, J.R., van de Sande, K.E., Gevers, T., Smeulders, A.W., 2013. Selective search for object recognition. Int. J. Comput. Vis. 104, 154–171.
Zeiler, M.D., Fergus, R., 2014. Visualizing and understanding convolutional networks. In: European Conference on Computer Vision, pp. 818–833.
Zheng, B., Wang, X., Lederman, D., Tan, J., Gur, D., 2010. Computer-aided detection; the effect of training databases on detection of subtle breast masses. Acad. Radiol. 17, 1401–1408.