Deep Learning Based Automatic Classification of Breast Lesions in Automated 3D Breast Ultrasound

Tao Tan[1]*, Mohsen Ghafoorian[1], Lei Wang[2],
Albert Gubern-Mérida[1], Jan van Zelst[1], Francesco Ciompi[1],
Brigitte Wilczek[3], Ritse M. Mann[1], Bram Platel[1],
Wei Zhang[4], Clara I. Sánchez[1],
Bram van Ginneken[1], Nico Karssemeijer[1]
[1]Department of Radiology and Nuclear Medicine,
Radboud University Medical Center, Nijmegen, the Netherlands.
[2]Fraunhofer Mevis, Bremen, Germany.
[3]Radiology, Unilabs S:T Göran mammography-department, Stockholm, Sweden
[4]QView Medical Inc, Los Altos, CA, USA.

August 3, 2015

Abstract

Purpose: Deep convolutional neural networks are attracting enormous attention in many machine learning applications, specifically in visual recognition tasks such as image and video analysis. In medical image analysis, there is also a trend to apply such techniques on different applications. In this paper, we investigate the possibility to use deep learning for lesion classification in automated 3D breast ultrasound.

Methods: Patches cut from a breast lesion volume at different positions with different sizes are used as input of an ensemble of convolutional neural network classifiers. The outputs of these classifiers are combined as features in another classifier. We performed receiver operating characteristic (ROC) analysis based on deep learning features. We also compared the performance of the deep learning based system with a previously developed CAD system using hand crafted features.

Results: The area under the ROC curve (Az) using deep learning alone is 0.82 which is lower than previous developed human-crafted features (Az=0.89; p <0.001). However, combining deep learning features and human-crafted features, the Az value was increased to 0.91 (p=0.03).

Conclusions: Using deep learning is beneficial to the existing classification system.

Key Words: breast ultrasound, automated 3D breast ultrasound, breast cancers, deep learning, convolutional neural networks, computer-aided diagnosis

1 Introduction

Ultrasound is a valuable adjunct to mammography in breast screening. Berg et al.[1] showed that adding ultrasound to mammography, the diagnostic yield increased from 7.6 per 1000 women screened to 11.8. The increased detection

*Corresponding author. Address: Geert Grooteplein Zuid 10, 6525 GA Nijmegen, The Netherlands.

of breast cancers are predominantly small and node-negative. However, the positive predictive value (PPV) of biopsy recommendation decreased from 22.6% to 11.2%. Ultrasound stand-alone has a PPV of 8.9%. In a recent study[2], researchers found that ultrasound is superior to mammography for breast cancer screening in high-risk Chinese women, while the economic cost of using ultrasound is much lower than that with mammography. Kelly et al.[3] recently showed that combining an automated scanning ultrasound system, breast cancer detection sensitivity increased significantly when compared with mammography alone in dense breasts while recalls increased from 4.2% to 9.6% adding ultrasound to mammography. Ultrasound has its value on improving the sensitivity of the cancer detection. However, the drawback is induced high recall rate and low PPV value or unnecessary benign biopsies.

To help radiologists classify breast lesions and be more selective for biopsies, computer aided diagnosis (CAD) in ultrasound has been studied in the last decades. In 3D ultrasound, Sahiner et al.[4] developed computer algorithms to combine image features extracted from width-to-height ratio, posterior shadowing, and texture descriptors. The system was shown to improve reader performance in the task of classifying breast lesions. Tan et al.[5] incorporated coronal spiculation, which is a very specific sign for malignancy in automated 3D breast ultrasound (ABUS), in a CAD system and a observer study was conducted showing that with the use of this system, the performance of inexperienced readers is improved[6]. Moon et al.[7] combined speckle and morphological features developing a CAD system for automated lesion classification. Liu et al.[8] combined texture features including local binary patterns, gray level co-occurrence matrix texture features and Gabor filters with features designed from the breast imaging reporting and data system standardized lexicon for US in an automated diagnosis system.

Basically in the existing literature mentioned above, the methods share a common general procedure which is as follows: First, the region of interest is identified. In this step, most methods require an accurate lesion segmentation. Some texture based methods require a bounding box surrounding the lesion. Second, features extracted from the boundary, shape, posterior shadowing, spiculation or retraction, and texture are extracted. Next, a classifier such as linear discriminant analysis, support vector machines, artificial neural network, etc is trained with labeled data. These steps make the backbone of a CAD system and the mentioned hand crafted features play a key role in the classification system. An important question now is can we design or find out more effective features that still contribute to a better detection for the CAD system. Researchers have developed various algorithms to extract features from regions of interest to differentiate malignant lesions from benign lesions and use supervised learning methods to train a classifier to predict the class or the likelihood of malignancy for new lesions.

Deep learning[9] has had many success stories in the past few years[10]. Convolutional neural networks (CNN)[11;12] are the most famous and successful variant of deep neural networks applied to visual recognition tasks in a supervised manner. So far there is no published work on application of CNNs or any other form of deep learning on computer-aided diagnosis for breast lesions. A challenge for this application is limitation of labeled samples of malignant and benign lesions. In this paper, we investigate the possibility of training a CNN for classification of breast lesions in automated 3D breast ultrasound and the possible contribution of a trained CNN to an existing CAD system.

2 Methods

2.1 Convolution neural networks

It has been shown[11] that CNNs are extremely powerful for automated classification of 2D input images. CNNs contain several layers of convolutions, each one optionally followed by pooling as a non-linear down-sampling functionality. After several layers of convolution and pooling, the intermediate outputs are usually fully connected to a multilayer perceptron neural network. The main advantage of CNNs is that they do not rely on hand crafting complicated features for characterization. Useful discriminative features for the classification task are automatically learned as parameters of the convolution filters. As two disadvantages of CNNs, they are usually computationally expensive and often demand a large amount of data for a decent training.

2.2 CNN training

2.2.1 Input patch preparation

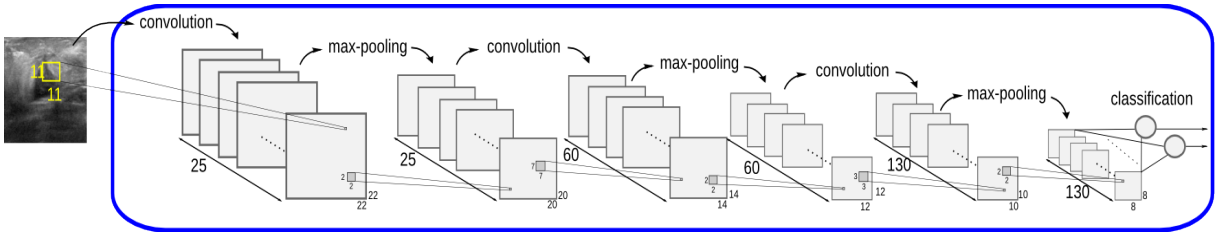
As CNNs expect fixed-sized patches as their input images, we first need to create fixed-sized patches from our dataset for training. Previously, we had applied a spiral scanning based dynamic programming method[13] to segment breast lesions in ultrasound. We used this method to segment breast lesions in ABUS. Suppose we want to train a CNN on 2D coronal plane at the center of the lesion. We computed the effective diameter d according to the segmentation size on this 2D slice. From the center of the segmentation, we sample a 2D patch of size $kd \times kd$, where k is used to control how much contextual or surrounding tissue information can be included in the patch. Given a fixed k for lesions with different sizes, the generated patch size would be different. To obtain fixed-sized patches, original patches are upsampled or downsampled to a fixed size (32×32).

2.2.2 Data augmentation and balancing

Due to hight number of weights that are to be learned and the natural complexity of CNNs, they are very prone for overfitting. A very common strategy to prevent this, is to use label-preserving transformations for data augmentation[14]. In particular we use translation, reflection and rotation to increment the number of positive patches. Since imbalanced datasets usually seriously deteriorate the performance of the CNN classifier, we create a class-balanced training dataset by selecting equal number of negative cases compared to the number of positive samples after augmentation.

2.2.3 CNN architecture and learning parameters

Our CNN consists of three layers: The First layer uses 25 filters and each filter has a size of 11×11. The convolution is followed by a max-pooling with a pool shape of 2×2 and stride step of 1×1. The second layer uses 60 filters of size of 7×7, followed by a max-pooling of shape 2×2 and stride step of 1×1. The last layer has 130 filters of shape 3×3 again followed a similar max-pooling structure. After these 3 layers, the output of the last layer is fully connected to an output layer of two nodes. After soft-max function, the output of these two nodes are used to optimize CNN parameters. A visualization of this network is depicted in Fig. 1. In this experiment, we use stochastic gradient descend learning algorithm with mini-batch size of 64 and learning rate of 0.001.



(a)

Figure 1: The architecture of CNN used in this work

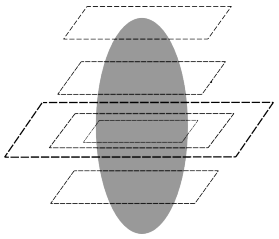
2.2.4 Parameter tuning

To prevent a bias in tuning the parameters of the system, we split the training data into two parts: training part and validating part. In this work, we used 70% of training data for the training and the remaining 30% for validating. Parameters are updated while using the training part for training and the error rates of the validation part are recorded so that the parameters set corresponding to the lowest validation error is used as the best classifier set of

parameters.

2.3 multi-patch classification

To be able to apply CNN on 3D volumes, we decompose a 3D volume into a series of 2D patches, which are inputs to CNN. In order to extract useful patches for the diagnosis, we take patches from three orthogonal planes, namely transversal, coronal and sagittal planes as radiologists often inspect lesions in these three planes. To fuse information from different planes, in this paper, we applied an ensemble of classifiers obtained from these patches. For each 2D view at a specific position, we trained a CNN classifier. Then we applied this CNN classifier on testing data. After applying different CNN classifiers for 2D patches from different locations, we combine the resulting likelihoods into a new likelihood indicating the malignancy of the 3D volume.



(a)

Figure 2: Patches cut in coronal plane

To select 2D patches to be representative from the 3D volume, at the center of the lesion, we use transversal, coronal and sagittal planes. To add context information, each plane is cropped with parameter k being 1, 2 and 3. Depending on the size of the lesion, extra patches are obtained by translation at the center. To cover the whole 3D volume, we also include transversal plane at one third of the lesion position and two thirds of the lesion position in the direction of the transducer translation. We also include coronal plane at one third of the lesion position and two thirds of the lesion position in the direction of the depth. For sagittal, we collect two extra patches in a similar way. In addition, as we observer that coronal spicluation is usually at the upper part of the lesion, and sometimes even at the top of the lesion, we add one patch at the top voxel of the lesion. For slices that are not the center of the 3D volume, we crop them with a size of 2.5 as the effective radius of the lesion. Fig. 2 demonstrates patches extracted from a lesion in coronal plane.

2.4 Previous human-crafted CAD system features

Previously, we had developed a CAD system combining region and texture features . Eleven features were extracted from ABUS volumes for discriminating between malignant and benign masses that included intensity variance, entropy, intensity average, margin contrast, volumetric height-to-width ratio, sphericity, compactness, posterior acoustic behavior and spiculation. Most of these features are mathematical descriptions of the features listed in the breast imaging reporting and data system (BI-RADS) standardized lexicon for US.

The texture features includes local binary patterns (LBP), gray level co-occurrence matrix (GLCM) based features and Gabor filters computed from 2D cross sections at the center of the lesion.

The number of computed texture features outnumbered the region features and also considerably exceeded the number of samples we had in our dataset. To deal with this issue we trained a separate classifier on each of our texture feature groups. These feature oriented classifiers (FOC) each combined the information of all texture features in a group into a single likelihood value. These likelihood values in turn were added to the 11 features of the existing CAD system. Therefore the previous CAD system uses 11 region features and 3 likelihood features from the three mentioned texture categories.

2.5 Experiments

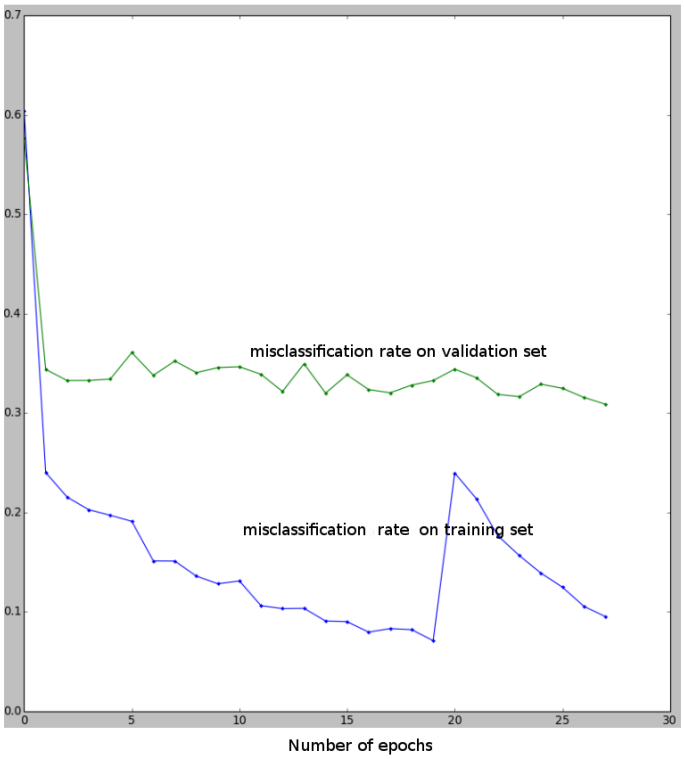
For this study we collected ABUS volumes from different institutes. We used two types of ABUS systems including SomoVu automated 3D breast ultrasound system developed by U-systems (Sunnyvale, CA, USA [1]) and ACUSON S2000 automated breast volume scanning system developed by Siemens (Mountain View, CA, USA). In total, we include 643 patients. For training and testing of our proposed method, we randomly split the data into two datasets: a training set (321 patients with 295 cancers and 500 benign lesions) and a testing set (322 patients, 366 cancers and 520 benign lesions).

The training set was split into a new training set and a validation set used for training of CNN classifiers. As likelihoods from different CNN classifiers for different types of patches form a feature vector, we trained a separate support vector machine (SVM) classifier to combine these features. Since upsampling and downsampling is performed when generating patches, we lose the information about the size of lesions. Therefore, we append the lesion size to the feature vector. We performed a 10-fold cross validation. For each run of the cross-validation, the

[1] On November 9, 2012, the healthcare division of General Electric-GE, announced the acquisition of U-Systems.

parameters of the SVM classifier is optimized using the training folds with a nested cross-validation. To compare to the previous CAD system, we also run a 10-fold cross-validation using features from the previous CAD system and we also combined deep learning features with existing features.

3 Results



(a)

Figure 3: misclassificiation rate of training and validation sets from saggittal plane at the center in the training epochs

Table 1 shows AUC values using different types of features extracted from the ABUS dataset. The AUC values of different features or combinations varied from 0.60 to 0.91. The deep learning features (Az=0.84) do not outperform previous human-crafted features (Az=0.89). However, by combining deep learning features and human-crafted features, the Az value increases to 0.91 (p=0.03).

Fig. 3 shows the misclassification rate of the training and validation sets during training. It can be seen that after the first epoch, the classification error dramatically decreases and afterwards, it decreases very slowly.

Checking each categories of deep learning features, likelihoods from transversal, coronal and sagittal lead to

a Az value of 0.77, 0.74 and 0.80 respectively. The likelihood from top coronal slice has a Az value of 0.60. A comparison of the ROC curves of the CAD system before and after adding deep learning features are shown in Fig.4.

Table 1: Performance of CAD system using different sets of features.

Feature(s)	number of features	AUC
human-crafted features	14	0.89
deep learning features	16	0.82
deep learning features + size	17	0.82
human-crafted + deep learning features	30	0.91
deep learning top coronal slice features	1	0.60
deep learning coronal features	6	0.74
deep learning transversal features	5	0.77
deep learning sagittal features	5	0.80

4 Conclusion and Discussion

In this work, we applied CNNs to the task of automated classification between benign and malignant lesions in automated 3D breast ultrasound. The input of the CNN classifier is a patch with a fixed size. CNNs automatically extract useful information via a large number of convolution filters and output a likelihood indicating the malignancy of a lesion. We used an SVM to combine CNN likelihoods targeting different areas of a lesion. The combination lead to an area under the ROC curve (Az) of 0.84. Although the discriminative power is lower than previous hand crafted features (Az=0.89; p <0.001), adding CNN likelihood as features to the existing features, the Az value was increased to 0.91 (p=0.03).

In this work, we decompose the 3D ABUS volume to a number of patches to cover 3D information and extract patches at the center of the segmentation on transversal, coronal and sagittal planes. Along the orthogonal axis through the center, we also take the patches at one-third and two-third position (don't know how say it better). Furthermore, as we observe that spiculation (associated to malignancy) also can be visible on the top of the lesion, we extract top coronal slice patch. We found out that the CNN from patch at top coronal slice lead to an Az of 0.56. Although the discrimination power is low, but there is still diagnostic value from the top slice. Among CNN likelihood features from different planes, likelihood from sagittal planes has the highest Az value which is surprising.

During the training, we did data argumentation, for example, for patches took from the center of the lesion, we

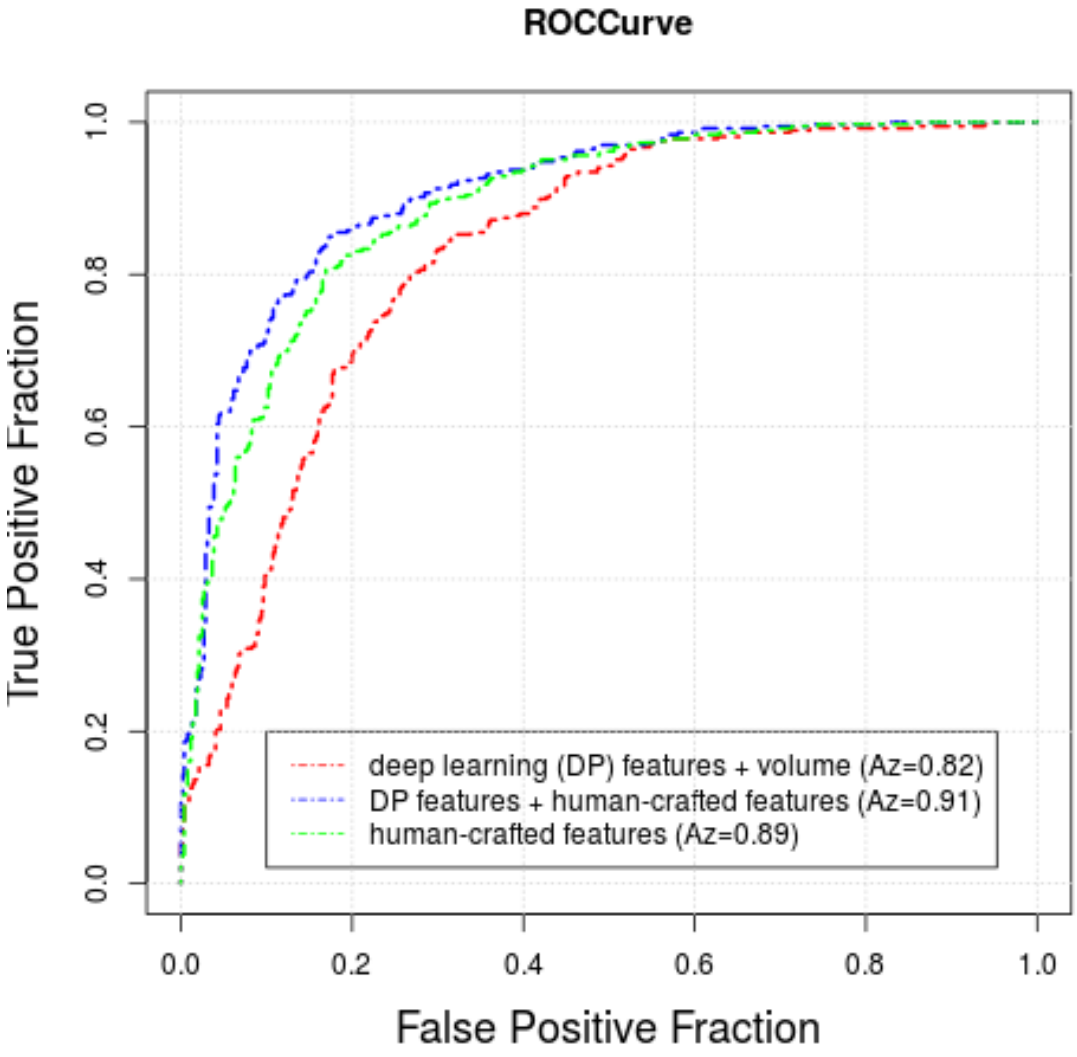


Figure 4: ROC curves with different feature sets.

also translate the center position by one or more voxel according to the size of lesion. Moreover, we rotate patches to have more training samples. With and without data argumentation, taking transversal plane as an example, the misclassification rate decrease from 0.28 to 0.25 after data augmentation. In other words, the more training data, the more discriminative the trained CNN is.

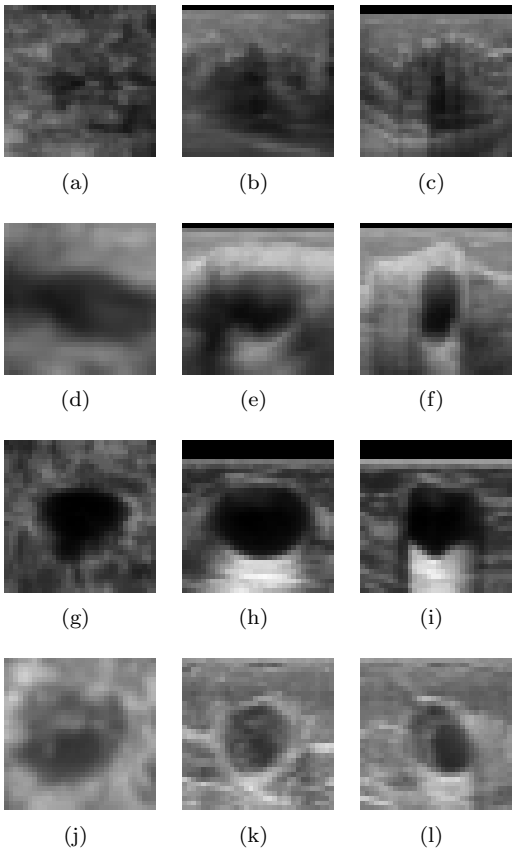
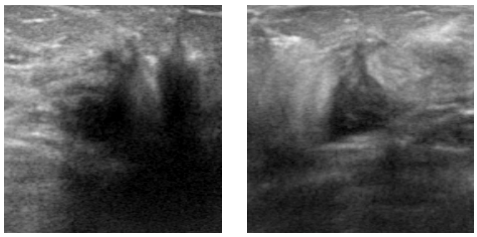


Figure 5: Examples of true positive(first row), false positive(second row), true negative(third row) and false negative(last row) patches from coronal (first column), transversal(second column) and sagittal plane (third column)

In many applications, deep learning approach already outperformed traditional hand crafted feature approach. However in our case, the performance is still worse than our traditional approach. One potential reason is that we do not have a huge amount of data for training CNN classifiers even after data augmentation. The other reason is that training a CNN is time consuming, and it is depending on the size of the patch. We only use a single scale for the patches and we downsample and upsample the original patches to a fixed size. During sampling, artificial structures might be created which could confuse the classifiers. An alternative approach is to choose the patch as big as the largest the lesion size in our dataset. However training a CNN using large patches would demand a lot of computation power and require a huge training data. The last reason might be that in ABUS, spiculation features

is very specific for cancers and it is mostly in coronal planes at upper part of the lesion. During patch selection, we only have three patch for upper part of coronal plane which might not fully cover the spiculation patterns of the cancers.

Fig .5 shows examples of true positive, false positive, true negative and false negative. It should also be noted that in ABUS imaging, due to multiple views with different compression directions, the same lesion may appear quite differently. Fig .6 shows the same benign lesion in medial and lateral view. The likelihoods computed from our CADx system are quite different because of the different appearances in these views. As a future work, the information from multiple views can be fused to make the classification more robust.



(a) (b)

Figure 6: A benign lesion from medial view (a) with a likelihood of 0.85 and lateral view(b) with a likelihood of 0.25.

References

[1] W. Berg, J. Blume, J. Cormack, E. Mendelson, D. Lehrer, M. Böhm-Vélez, E. Pisano, R. Jong, W. Evans, M. Morton, M. Mahoney, L. Larsen, R. Barr, D. Farria, H. Marques, K. Boparai, and A. . I. , "Combined screening with ultrasound and mammography vs mammography alone in women at elevated risk of breast cancer," **299**, 2151–2163 (2008).

[2] S. Shen, Y. Zhou, Y. Xu, B. Zhang, X. Duan, R. Huang, B. Li, Y. Shi, Z. Shao, H. Liao, J. Jiang, N. Shen, J. Zhang, C. Yu, H. Jiang, S. Li, S. Han, J. Ma, and Q. Sun, "A multi-centre randomised trial comparing ultrasound vs mammography for screening breast cancer in high-risk chinese women," (2015).

[3] K. Kelly, J. Dean, W. Comulada, and S. Lee, "Breast cancer detection using automated whole breast ultrasound and mammography in radiographically dense breasts," **20**, 734–742 (2010).

[4] B. Sahiner, H. Chan, M. Roubidoux, L. Hadjiiski, M. Helvie, C. Paramagul, J. Bailey, A. Nees, and C. Blane,

"Malignant and benign breast masses on 3d us volumetric images: effect of computer-aided diagnosis on radiologist accuracy," **242**, 716–724 (2007).

[5] T. Tan, B. Platel, H. Huisman, C. I. Sánchez, R. Mus, and N. Karssemeijer, "Computer aided lesion diagnosis in automated 3D breast ultrasound using coronal spiculation," **31**, 1034–1042 (2012).

[6] T. Tan, B. Platel, T. Twellmann, G. van Schie, R. Mus, A. Grivegnée, R. M. Mann, and N. Karssemeijer, "Evaluation of the effect ofcomputer-aided classification ofbenign and malignant lesions onreader performance in automated three-dimensional breast ultrasound," **20**, 1381–1388 (2013).

[7] W. K. Moon, C.-M. Lo, J. M. Chang, C.-S. Huang, J.-H. Chen, and R.-F. Chang, "Computer-aided classification of breast masses using speckle features of automated breast ultrasound images," **39**, 6465–6473 (2012).

[8] H. Liu, T. Tan, J. van Zelst, R. Mann, N. Karssemeijer, and B. Platel, "Incorporating texture features in a computer-aided breast lesion diagnosis system for automated three-dimensional breast ultrasound," **1**, 024501–024501 (2014).

[9] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," **521**, 436–444 (2015).

[10] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Networks* **61**, 85–117 (2015).

[11] B. B. Le Cun, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Handwritten digit recognition with a back-propagation network," in *Advances in neural information processing systems*, Citeseer (1990).

[12] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE* **86**(11), 2278–2324 (1998).

[13] T. Tan, B. Platel, R. Mus, and N. Karssemeijer, "Detection of breast cancer in automated 3D breast ultrasound," **8315**, 831505–1–831505–8 (2012).

[14] P. Y. Simard, D. Steinkraus, and J. C. Platt, "Best practices for convolutional neural networks applied to visual document analysis," in *Document Analysis and Recognition*, 958–963 (2003).

List of Figures

List of Tables