Sergei V. Fotin, Yin Yin, Hrishikesh Haldankar, Jeffrey W. Hoffmeister and Senthil Periaswamy, "Detection of soft tissue densities from digital breast tomosynthesis: comparison of conventional and deep learning approaches." Medical Imaging 2016: Computer-Aided Diagnosis, edited by Georgia D. Tourassi, Samuel G. Armato III, Proc. of SPIE Vol. 9785, 97850X

http://dx.doi.org/10.1117/12.2217045

Detection of soft tissue densities from digital breast tomosynthesis: comparison of conventional and deep learning approaches

Sergei V. Fotin, Yin Yin, Hrishikesh Haldankar, Jeffrey W. Hoffmeister, Senthil Periaswamy

iCAD Inc., 98 Spit Brook Road, Suite 100, Nashua, NH 03062, USA

ABSTRACT

Computer-aided detection (CAD) has been used in screening mammography for many years and is likely to be utilized for digital breast tomosynthesis (DBT). Higher detection performance is desirable as it may have an impact on radiologist's decisions and clinical outcomes. Recently the algorithms based on deep convolutional architectures have been shown to achieve state of the art performance in object classification and detection. Similarly, we trained a deep convolutional neural network directly on patches sampled from two-dimensional mammography and reconstructed DBT volumes and compared its performance to a conventional CAD algorithm that is based on computation and classification of hand-engineered features. The detection performance was evaluated on the independent test set of 344 DBT reconstructions (GE SenoClaire 3D, iterative reconstruction algorithm) containing 328 suspicious and 115 malignant soft tissue densities including masses and architectural distortions. Detection sensitivity was measured on a region of interest (ROI) basis at the rate of five detection marks per volume. Moving from conventional to deep learning approach resulted in increase of ROI sensitivity from 0.832 ± 0.040 to 0.893 ± 0.033 for suspicious ROIs; and from 0.852 ± 0.065 to 0.930 ± 0.046 for malignant ROIs. These results indicate the high utility of deep feature learning in the analysis of DBT data and high potential of the method for broader medical image analysis tasks.

Keywords: Computer-aided detection (CAD), deep learning, convolutional neural networks (CNN), digital breast tomosynthesis (DBT), digital mammography

1. INTRODUCTION

Digital breast tomosynthesis (DBT)[1] is a relatively new imaging modality in which systems image a breast by moving an X-ray source and exposing the breast to radiation from multiple angles, thus acquiring high resolution, planar digital projections.[2] Then, given the projection data, a 3D DBT volume is typically created using one of the reconstruction algorithms.[3] Such 3D images enable physicians to visualize important mammographic structures at specified reconstruction heights reducing the effect of the structures obscuring lesion visibility on a typical 2D mammography image. Multiple clinical studies confirmed an increased value of adding DBT to conventional 2D digital mammography by showing improved sensitivity and/or specificity of radiological evaluations.[4–6]

Depending on the breast size and reconstruction interval, the number of reconstructed slices in a DBT volume may be well above one hundred. The increase in the cost of manual interpretation creates a demand for the fast and accurate detection algorithms. While a lot of up to date research was focused on CAD for 2D mammography, very few assessed CAD solutions for DBT either for detection of soft tissue densities[7–13] or calcification clusters.[14–18]

A naive approach of applying CAD for DBT would be the adaptation of the existing solutions for 2D mammography. Running CAD on reconstructed DBT volumes in a slice-by-slice fashion followed by grouping of clustering of neighboring detections may result in a performance comparable to 2D mammography. The less explored path is to develop a CAD algorithm that runs directly on low-dose raw projection images. To compensate for the absent high frequency details in raw images, one would need to consider all projections in conjunction making such analysis somewhat equivalent to detection from reconstructed volumes.

Send correspondence to Sergei V. Fotin to sergei.fotin(at)gmail(dot)com

A conventional 2D mammography CAD scheme consists of several steps that may include image preprocessing, breast segmentation, candidate generation, feature extraction and classification. In a certain scenario, one may include a postprocessing step that would combine information from multiple views of the breast to produce more accurate case level decisions. To develop high-performance system, the engineers must spend a lot of time designing, developing and validating individual features that would ultimately increase the performance of the system.

Recently there has been a lot of progress in alternative end-to-end learning techniques, notably the work of Krizhevsky et al.[19] on deep convolutional neural networks (CNN), where the learning and classification is performed using raw pixel data. His team applied the ideas of LeCun et al.[20] to the problem of large-scale image classification. Subsequently solutions based on this architecture were able to achieve state-of-the-art results in object detection from natural images.[21] It is important to emphasize that convolutional neural networks have been used in the past for mammographic image analysis,[22–24] however the deep architectures were not extensively explored due to considerable computation costs.

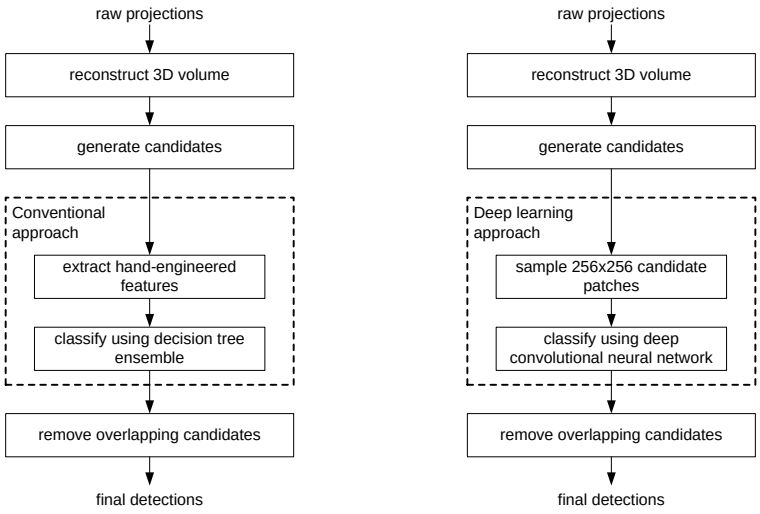
The purpose of this work is to evaluate conventional and deep convolutional neural network approaches in detection and recognition of cancers from 3D reconstructed DBT volumes.

2. METHOD

Our conventional CAD approach was based on an algorithm, originally designed to run on 2D mammography images and modified to run on a slice-by-slice basis on the reconstructed DBT. The algorithm consisted of a high sensitivity multiscale candidate generator targeting two soft tissue density abnormality types: masses and architectural distortions. Each candidate was defined by its coordinates within the reconstructed DBT volume or a 2D mammography image and approximate size. A simple algorithm based on conditional random field inference was used to segment the mass candidates. The candidates were then passed to a feature extractor that was computing over 300 carefully tuned features including multiscale contrast, histogram, gradient, texture, shape and topology descriptors. The final classifier was trained on a combination of multi-vendor 2D mammography and 3D DBT data. We selected an ensemble of boosted decision trees as the classifier in the conventional setup as it had been internally demonstrated to achieve highest classification performance and robust vendor independence among all explored classification methods. In the postprocessing step, we eliminated majority of overlapping detections leaving only the ones with the locally maximal classification score. It is possible, however, that some of the detections would still remain clustered together with little or no overlap. The sketch of the conventional CAD approach is shown in Figure 1(a).

The difference between the deep learning based and the conventional approaches is in the feature extraction and classification steps that were substituted by a deep CNN operating directly on sampled image patches as illustrated in Figure 1(b). Again, the detection pipeline is started with the same candidate generator producing mass and architectural distortion candidates. Then, for each candidate a square patch was constructed around it with the side of the square equal to three estimated diameters of the candidate. Such sizing allowed us to capture full candidate extent plus some surrounding context. Each square patch was then sampled from the image to have the resolution of 256 by 256 pixels by means of bilinear interpolation. Finally, the brightness of these patches was linearly scaled to have zero mean and unit standard deviation.

We used BVLC Caffe deep learning framework[25] to train a deep CNN with the architecture almost identical to "AlexNet"[19] that consists of multiple subsequent convolutional layers with maximal pooling and linear rectification in between. The scaled image patches were then used as the inputs and the top softmax unit was set to discriminate between four classes of candidates: positive and negative masses, positive and negative architectural distortions. All four classes were randomly sampled from the set of candidates produced by the candidate generator, while the set of positive candidates was augmented by random shifting, rotation and scaling. We used default network weights initialization and training parameters as supplied by the framework. To process the sampled patch data, we substituted default three channel RGB input to a single grayscale floating point input. The learning algorithm was set up to minimize multinomial logistic loss of the softmax with respect to four class assignments using batch gradient descent. At test time, positive and negative classification was done by comparing the total probabilities of positive and negative assignments.



(a) conventional approach (b) deep learning approach

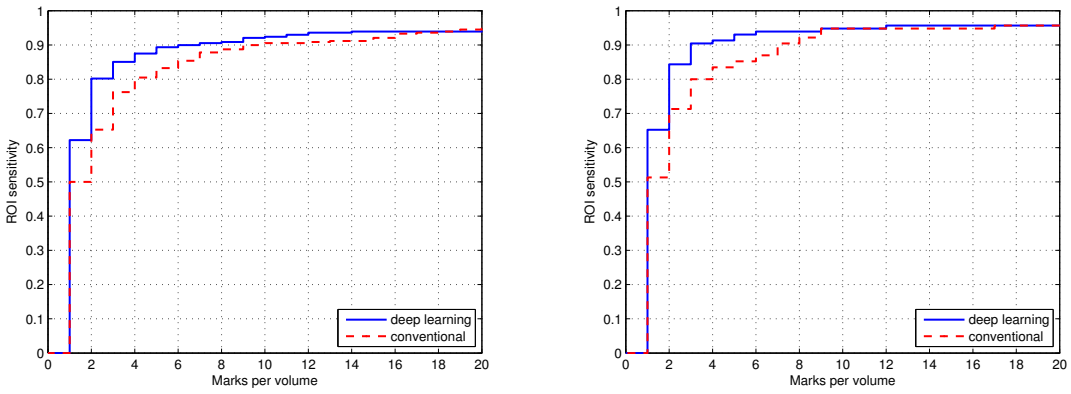
Figure 1. Algorithm scheme for two CAD configurations: (a) feature extraction and classification is performed in conventional way; and (b) sampled candidate patches are classified by deep convolutional neural network.

The remaining CAD step of removing overlapping detections was exactly the same as in the conventional approach.

3. EXPERIMENT

Because of the limited availability of malignancy-proven cancer cases in our database, we used all suspicious findings that had been annotated by a radiologist, including cancers and suspicious lesions with non-negative BI-RADS[26] ratings. Each annotation contained a hand-drawn contour of the lesion on a reconstructed slice with the sharpest lesion appearance. The contour was supplemented with top and bottom delimiting points indicating the visible extent of the lesion along z-axis. We assembled a training set containing 1864 suspicious soft tissue density lesions from 2D mammography and 339 lesions from DBT. Approximately a third of these lesions were proven to be malignant. The 2D mammography subset contained digitized film, CR and DR mammography images from various vendors. The DBT subset contained exclusively GE SenoClaire 3D DBT volumes. To train a classifier (either conventional or CNN), for each lesion we identified at most 5 candidates having the largest overlap with the radiolodist's annotation and used them as positive examples. For each true positive example we randomly sampled 10 negative examples from the output of the candidate generator.

Evaluation of both detection schemes was done on an independent test set of 344 reconstructed DBT volumes (GE SenoClaire 3D) containing 328 suspicious lesions among which 115 were proven malignancies. Every detection mark was counted as a true positive if it had an overlap with the ground truth contour and lied within the top and bottom delimiting points. A vast majority of the cases that were used for evaluation contained only a single DBT reconstruction per breast; therefore we reported only the sensitivity on region of interest (ROI) level which is not directly comparable to the case-based or image-based sensitivity obtained from two-view mammography. We measured sensitivity as a fraction of true positive lesion ROIs to the total number of lesion ROIs in the test dataset. After the classification and overlap elimination steps had been done, we ranked all remaining detections by their classification score and produced multiple operating points by changing the rank threshold within integer range between 0 and 20. We prefer the reporting of marks per volume on the x-axis to the rate of false positives, as this gives us more consistent measurements of sensitivity and independence of dataset characteristics (such as ratio of cancer cases to normal cases) across different datasets. Moreover, at the operating point of 5 marks per image we measured the 95% confidence intervals for the mean sensitivity, average false positive rate per evaluation set volume and the z-distance between the ground truth contour annotation and the corresponding detection mark.



(a) Evaluation set of 328 suspicious lesions (b) Evaluation set of 115 malignant lesions

Figure 2. Comparison of conventional and deep learning approaches as a trade-off between region of interest (ROI) sensitivity and the number of detection marks generated per DBT volume. Performance is evaluated on (a) 328 suspicious lesions and (b) 115 proven malignant lesions.

Table 1. Comparison of conventional and deep learning approaches at an operating point of 5 detection marks per image.

Evaluation set	328 suspicious lesions		115 malignant lesions	
Detection approach	Conventional	Deep learning	Conventional	Deep learning
Mean ROI Sensitivity	0.832 ± 0.040	0.893 ± 0.033	0.852 ± 0.065	0.930 ± 0.046
Detection marks per volume	5	5	5	5
Mean false positives per volume	3.50	3.25	3.28	3.11
Mean vertical offset, mm	2.79	2.59	2.57	1.95

In addition to performance comparison for two approaches, we also constructed a learning curve to understand and estimate behavior of the deep convolutional neural network depending on amount of available training data. We obtained the measurements of the CNN's test accuracy in the scenarios when only a fraction (0.1, 0.2 and 0.5) of the whole training data is available. For each fraction value, we randomly sampled the training data three times and measured test accuracy as the mean value and 95% confidence intervals for the mean.

4. RESULTS AND DISCUSSION

The candidate generator that has been used to start off the experiment provided the baseline sensitivity of 1.0 for both malignant and suspicious lesions at an average rate of 1935 marks per volume. The 94.2% of the true positive detections had an average Jaccard index over 0.5 when comparing to the ground truth contouring, which indicates the quality of localization and segmentation provided by candidate generator. The deep CNN achieved classification accuracy of 0.8640. The comparison of detection performance curves for two explored approaches is shown in Figure 2. Detailed comparison of detection performance at 5 detection marks per image is given in Table 1. As we can see, the deep learning approach clearly outperforms the conventional approach in terms of sensitivity in both lesion categories: suspicious (0.832 vs. 0.893) and malignant (0.852 vs. 0.930). Not only does it provide much higher sensitivity at the same detection mark rate, but it also results in detections that are closer to the "sharpest" ground truth slice along z-axis. The number of false positives is also reduced slightly for the deep learning approach indicating that the detections tend to cluster together near the regions of interest.

These results are particularly interesting in the light of the amount of time needed to engineer deep learning solution which was measured in several weeks as opposed to months (or years) of conventional feature design and development. The explored deep architecture allows us to transfer the concepts learned from the analysis of mammography training data to the detection of soft tissue densities from digital breast tomosynthesis. Moreover, there is a potential of applying deep CNNs in a sliding window fashion over the entire image that would eliminate the need of conventional detection steps such as candidate generation and segmentation.

The learning curve in Figure 3 shows that the test accuracy increases with the amount of the training data and still have the capacity to grow if the deep CNN's training algorithm received more data. It is reasonable

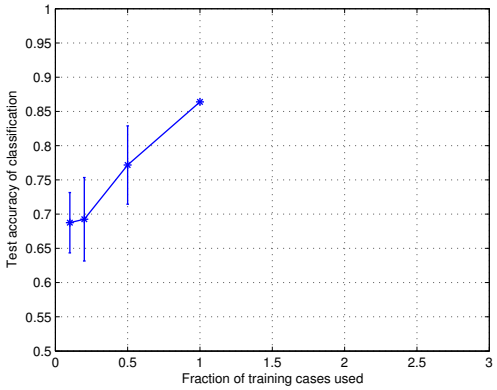


Figure 3. Learning curve displaying the effect of training set size on the candidate classification performance. Fraction value of 1 on the x-axis corresponds to using all 1864 suspicious soft tissue density lesions from 2D mammography and 339 lesions from DBT for deep CNN training. At each of the lower fractions of 0.1, 0.2 and 0.5, the training set was randomly subsampled three times and resultant mean test accuracy value was plotted along with the 95% confidence intervals.

to assume that the increase in the CNN's accuracy would lead to the increase of the detection performance, however we did not perform such measurements within this work.

Among the drawbacks of the deep learning comparing to the conventional approach is that the runtime during the test stage is substantially longer. On the dual Xeon 2.40 GHz workstation and custom Caffe modification it takes approximately 1 minute to classify 1000 candidate patches in CPU mode, while conventional approach takes about 20 seconds. Moving the computation to specialized hardware or GPU will help to reduce the computation time at the expense of the system cost.

5. CONCLUSION

In comparison of conventional approach to deep learning we observed high utility of the latter in the analysis of digital breast tomosynthesis data and high potential of the method for broader medical image analysis tasks. Deep feature learning is a new technology enabling rapid development of computer-aided detection systems superior to the conventional, old generation approaches.

REFERENCES

[1] Niklason, L. T., Christian, B. T., Niklason, L. E., Kopans, D. B., Castleberry, D. E., Opsahl-Ong, B., Landberg, C. E., Slanetz, P. J., Giardino, A. A., Moore, R., et al., "Digital tomosynthesis in breast imaging," *Radiology* **205**(2), 399–406 (1997).

[2] Sechopoulos, I., "A review of breast tomosynthesis. Part I. The image acquisition process," *Medical physics* **40**(1), 014301 (2013).

[3] Sechopoulos, I., "A review of breast tomosynthesis. Part II. Image reconstruction, processing and analysis, and advanced applications," *Medical physics* **40**(1), 014302 (2013).

[4] Skaane, P., Bandos, A. I., Gullien, R., Eben, E. B., Ekseth, U., Haakenaasen, U., Izadi, M., Jebsen, I. N., Jahr, G., Krager, M., et al., "Comparison of digital mammography alone and digital mammography plus tomosynthesis in a population-based screening program," *Radiology* **267**(1), 47–56 (2013).

[5] Ciatto, S., Houssami, N., Bernardi, D., Caumo, F., Pellegrini, M., Brunelli, S., Tuttobene, P., Bricolo, P., Fantò, C., et al., "Integration of 3D digital mammography with tomosynthesis for population breast-cancer screening (STORM): a prospective comparison study," *The lancet oncology* **14**(7), 583–589 (2013).

[6] Friedewald, S. M., Rafferty, E. A., Rose, S. L., Durand, M. A., Plecha, D. M., Greenberg, J. S., Hayes, M. K., Copit, D. S., Carlson, K. L., Cink, T. M., et al., "Breast cancer screening using tomosynthesis in combination with digital mammography," *JAMA* **311**(24), 2499–2507 (2014).

[7] Reiser, I. and Nishikawa, R., "Computerized Mass Detection for Digital Breast Tomosynthesis," *Recent Advances in Breast Imaging, Mammography, And Computer-Aided Diagnosis of Breast Cancer* **155**, 409 (2006).

[8] Chan, H.-P., Wei, J., Zhang, Y., Helvie, M. A., Moore, R. H., Sahiner, B., Hadjiiski, L., and Kopans, D. B., "Computer-aided detection of masses in digital tomosynthesis mammography: Comparison of three approaches," *Medical physics* **35**, 4087 (2008).

[9] Singh, S., Tourassi, G. D., Baker, J. A., Samei, E., and Lo, J. Y., "Automated breast mass detection in 3D reconstructed tomosynthesis volumes: A featureless approach," *Medical physics* **35**, 3626 (2008).

[10] Palma, G., Muller, S., Bloch, I., and Iordache, R., "Fast detection of convergence areas in digital breast tomosynthesis," in [*Biomedical Imaging: From Nano to Macro, 2009. ISBI'09. IEEE International Symposium on*], 847–850, IEEE (2009).

[11] Palma, G., Bloch, I., and Muller, S., "Spiculated lesions and architectural distortions detection in digital breast tomosynthesis datasets," in [*Digital Mammography*], 712–719, Springer (2010).

[12] Mazurowski, M. A., Lo, J. Y., Harrawood, B. P., and Tourassi, G. D., "Mutual information-based template matching scheme for detection of breast masses: From mammography to digital breast tomosynthesis," *Journal of biomedical informatics* **44**(5), 815–823 (2011).

[13] van Schie, G., Wallis, M. G., Leifland, K., Danielsson, M., and Karssemeijer, N., "Mass detection in reconstructed digital breast tomosynthesis volumes with a computer-aided detection system trained on 2D mammograms," *Medical physics* **40**, 041902 (2013).

[14] Peters, G., Muller, S., Bernard, S., Iordache, R., Wheeler, F., and Bloch, I., "Reconstruction-independent 3D CAD for calcification detection in digital breast tomosynthesis using fuzzy particles," in [*Progress in Pattern Recognition, Image Analysis and Applications*], 400–408, Springer (2005).

[15] Park, S. C., Zheng, B., Wang, X.-H., and Gur, D., "Applying a 2D based CAD scheme for detecting microcalcification clusters using digital breast tomosynthesis images: An assessment," in [*Medical Imaging*], 691507–691507, International Society for Optics and Photonics (2008).

[16] Bernard, S., Muller, S., and Onativia, J., "Computer-aided microcalcification detection on digital breast tomosynthesis data: A preliminary evaluation," in [*Digital Mammography*], 151–157, Springer (2008).

[17] Chan, H.-P., Wu, Y.-T., Sahiner, B., Wei, J., Helvie, M. A., Zhang, Y., Moore, R. H., Kopans, D. B., Hadjiiski, L., and Way, T., "Characterization of masses in digital breast tomosynthesis: Comparison of machine learning in projection views and reconstructed slices," *Medical physics* **37**, 3576 (2010).

[18] Spangler, M. L., Zuley, M. L., Sumkin, J. H., Abrams, G., Ganott, M. A., Hakim, C., Perrin, R., Chough, D. M., Shah, R., and Gur, D., "Detection and classification of calcifications on digital breast tomosynthesis and 2D digital mammography: a comparison," *American Journal of Roentgenology* **196**(2), 320–324 (2011).

[19] Krizhevsky, A., Sutskever, I., and Hinton, G. E., "Imagenet classification with deep convolutional neural networks," in [*Advances in neural information processing systems*], 1097–1105 (2012).

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

[21] Girshick, R., Donahue, J., Darrell, T., and Malik, J., "Rich feature hierarchies for accurate object detection and semantic segmentation," in [*Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*], 580–587, IEEE (2014).

[22] Zhang, W., Doi, K., Giger, M. L., Wu, Y., Nishikawa, R. M., and Schmidt, R. A., "Computerized detection of clustered microcalcifications in digital mammograms using a shift-invariant artificial neural network," *Medical Physics* **21**(4), 517–524 (1994).

[23] Sahiner, B., Chan, H.-P., Petrick, N., Wei, D., Helvie, M. A., Adler, D. D., and Goodsitt, M. M., "Classification of mass and normal breast tissue: a convolution neural network classifier with spatial domain and texture images," *Medical Imaging, IEEE Transactions on* **15**(5), 598–610 (1996).

[24] Lo, S.-C. B., Li, H., Wang, Y., Kinnard, L., and Freedman, M. T., "A multiple circular path convolution neural network system for detection of mammographic masses," *Medical Imaging, IEEE Transactions on* **21**(2), 150–158 (2002).

[25] Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., and Darrell, T., "Caffe: Convolutional architecture for fast feature embedding," *arXiv preprint arXiv:1408.5093* (2014).

[26] D'Orsi, C. J. et al., [*ACR BI-RADS Atlas: Breast Imaging Reporting and Data System*] (2013).