Automated Mass Detection in Mammograms using Cascaded Deep Learning and Random Forests

Neeraj Dhungel
ACVT, School of Computer Science
The University of Adelaide

Gustavo Carneiro
ACVT, School of Computer Science
The University of Adelaide

Andrew P. Bradley†
School of ITEE,
The University of Queensland

***Abstract*—Mass detection from mammograms plays a crucial role as a pre-processing stage for mass segmentation and classification. The detection of masses from mammograms is considered to be a challenging problem due to their large variation in shape, size, boundary and texture and also because of their low signal to noise ratio compared to the surrounding breast tissue. In this paper, we present a novel approach for detecting masses in mammograms using a cascade of deep learning and random forest classifiers. The first stage classifier consists of a multi-scale deep belief network that selects suspicious regions to be further processed by a two-level cascade of deep convolutional neural networks. The regions that survive this deep learning analysis are then processed by a two-level cascade of random forest classifiers that use morphological and texture features extracted from regions selected along the cascade. Finally, regions that survive the cascade of random forest classifiers are combined using connected component analysis to produce state-of-the-art results. We also show that the proposed cascade of deep learning and random forest classifiers are effective in the reduction of false positive regions, while maintaining a high true positive detection rate. We tested our mass detection system on two publicly available datasets: DDSM-BCRP and INbreast. The final mass detection produced by our approach achieves the best results on these publicly available datasets with a true positive rate of 0.96 ± 0.03 at 1.2 false positive per image on INbreast and true positive rate of 0.75 at 4.8 false positive per image on DDSM-BCRP.**

I. Introduction

Breast cancer is considered to be one of the most common cancers affecting women around the world. According to the World Cancer Report [1], breast cancer accounts for 22.9% of diagnosed cancers and 13.7% of cancer related deaths worldwide. Around 15% of the cancer related deaths affecting Australian women are due to breast cancer, accounting for the second highest death rate caused by any type of cancer [2]. Since there are various causes associated breast cancer (with some of them still unknown), the best tool to reduce the mortality rate is early detection using different imaging modalities such as x-ray (mammography), ultrasound and magnetic resonance imaging (MRI).

Mammography is a widely used imaging modality for screening breast cancer because it enables the detection of wide variety of suspicious lesions (e.g., masses and micro-clacifications). This is the first step in assessing the risk of having, or developing breast cancer. However, the current problem is that mass detection is mostly a manual process, so a significant number of breast masses can be missed or those which are detected turn out to be benign after biopsy [3], [4]. In fact, studies show a sensitivity of between 85%-90% in cancer detection [5] from mammography. In part, this happens because of the variability in shape, size, boundary and texture of masses [6], [3], and also because of their low signal-to-noise ratio compared to surrounding normal breast tissue, as depicted in Fig.1. Masses are visually characterized by medium gray to white regions in the breast area of mammograms, and their shapes are generally described as oval, irregular, or lobulated, with boundaries that can be circumscribed, obscured, ill-defined or spiculated [3], [4]. Detection sensitivity can be improved in clinical practice with the use of an automated mass detection system that acts as a "second" opinion [7], which can reduce to some extent, the current dependence on the radiologist's experience and workload [8]. Studies on the effectiveness of the use of computer-aided diagnosis (CAD) systems as second opinion systems show that they can indeed help junior radiologists by increasing their sensitivity from from 62% to 85% and experienced radiologists by increasing their sensitivity from 77% to 85% [9]. This shows the neccessity of developing CAD systems that are robust to false positive detection of lesions and at the same time effective in their true positive detections.

The main goal of this paper is to present a new approach for the detection of masses from mammograms that combines two of the most powerful machine learning techniques developed in last few years: deep learning and random forests (see Fig. 2). In our method, the first stage consists of a multi-scale deep belief network (m-DBN) cascade of classifiers [10] combined with a Gaussian mixture model (GMM) [11] classifier that selects a set of regions, representing candidate breast masses. This first stage is followed by a second stage, comprising a cascade of deep convolutional neural networks (CNN) [12], [13] that reduce the number of false postives, while keeping the vast majority of the true positive regions. Finally, we extract texture and morphological features from the remaining regions so they can be classified by a cascade of random forest classifiers [14]. The motivation behind the use of the cascade of random forest classifiers is that the performance of the deep learning (m-DBN and CNN) cascade of classifiers saturates after a certain number of stages. Hence, at this point,

†This work was partially supported by the Australian Research Council's Discovery Projects funding scheme (project DP140102794). Prof. Bradley is the recipient of an Australian Research Council Future Fellowship (FT110100623).

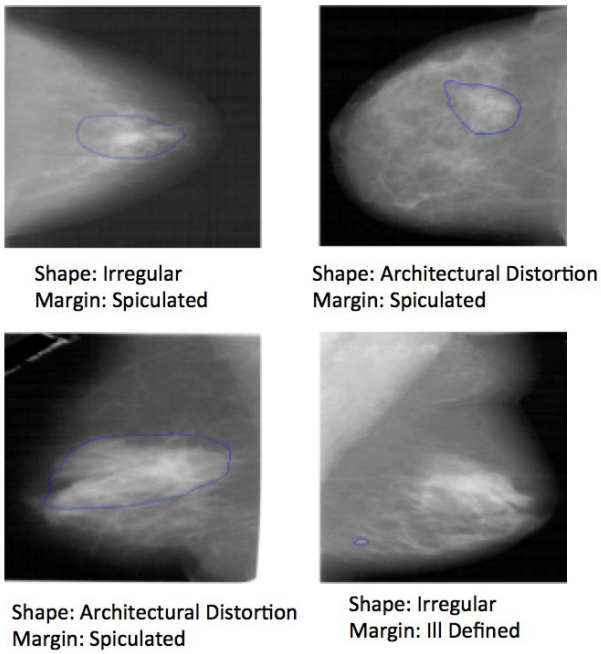

Fig. 1: Some categories of malignant masses from DDSM-BCRP dataset. The contour in blue indicates the suspected malignant masses annotated by experts.

we extract hand-designed domain specific features from the candidates regions selected by the deep learning cascade and apply the random forest (RF) cascade (note that we use RF for this final stage given its superior performance in the recently published study by Fernandez et al. [15]). We show that our methodology produces the best mass detection results to date on the public datasets DDSM-BCRP [16] and INbreast [17] using mammograms containing normal tissue, malignant and benign masses.

II. Literature Review

In spite of the development of numerous mass detection systems, they are still not widely used in clinical practice because they tend to generate a large number of false positives in a test mammographic image [18]. Bilateral image subtraction [19] which implements difference between the left and right breast to obtain the suspicious regions, was one of the earliest method employed for the detection of the masses in mammograms. Most of the other mass detection systems have a first stage that detects candidate regions using several filters, such as morphological, difference of Gaussian, and Laplacian of Gaussian filters [20], [21], [22], [23], [24], [25], [26], [27]. This stage is followed by a false positive removal step, using different types of classifiers, such as support vector machine, linear discriminant analysis or neural networks [20], [21], [22], [23], [24], [25], [26], [27]. One of the main drawbacks of such systems is that they can generate a large number of false positives, while missing a good proportion of true positives [3]. Another problem is that they are usually tested on private datasets or on selected subsets of DDSM [16] that are not publicly available, which makes direct comparisons difficult [28]. An important point to mention here is that researchers in this field have used these private datasets in spite of the availability of public datasets such as DDSM-BCRP [16] and INbreast [17]. The fact that DDSM-BCRP is not commonly used is because there is little consensus with its experimental setup and ground truth annotation [28]. INbreast [17] dataset is more recent and contains high quality images and ground-truth annotation, so it consists of an ideal test-bed for comparing mass detection systems. In addition to these problems, most of the current methods [20], [21], [22], [23] are tested only on mammograms that contain malignant masses rather than malignant, benign masses and completely normal mammograms, which biases the results.

Recently, deep learning methods, and in particular, deep convolutional neural nets (CNN), have produced state-of-the-art results in several computer vision problems. It has been shown that stochastic gradient descend via backpropagation [12] is an effective tool for training CNNs. The increasing depth of the network increases the performance of the system by an impressive margin on the ImageNet Challenge [29], [30] by using different types of regularisation techniques, such as dropout [13]. Having a very deep network is certainly useful in extracting features that are effective for classification, but at the expense of computation time and memory requirement. In this regard, cascaded CNNs can be seen as a way to maintain the depth of the network while reducing the computation hurdles. More recently, cascaded CNNs with sliding windows have been successfully applied in the face detection problem, producing the current state-of-the-art result [31], but its design was carefully planned to avoid a large run time complexity. Another successful approach using deep learning is the R-CNN [32], which generates a large pool of candidates and classifies them using CNNs [12], [32]. The main advantage of R-CNN [32] is that they use the concept of region recognition, which is computationally more efficient than the sliding window over the whole image. Our method is similar to R-CNN, but we use a cascade of R-CNNs with the aim of reducing the number of false positive detection. For object region proposal generation we use another class of deep learning method called deep belief nets (DBN) [10] on several image resolutions, which is called multi-scale DBN (m-DBN).

III. Methodology

Our mass detection system consists of four modules, as shown in Fig. 2. The first module combines a multi-scale deep belief network (m-DBN) [10], [6], shown in Fig. 3, with a Gausssian mixture model [11] (GMM) classifier for candidate generation. These candidates are the inputs to a second stage containing a cascade of two deep convolutional neural networks (CNN) [12], [13] that produce features that are used by a linear support vector machine (SVM) classifier (this combination of CNN applied to specific image regions and SVM is known as an R-CNN [32] in the computer vision

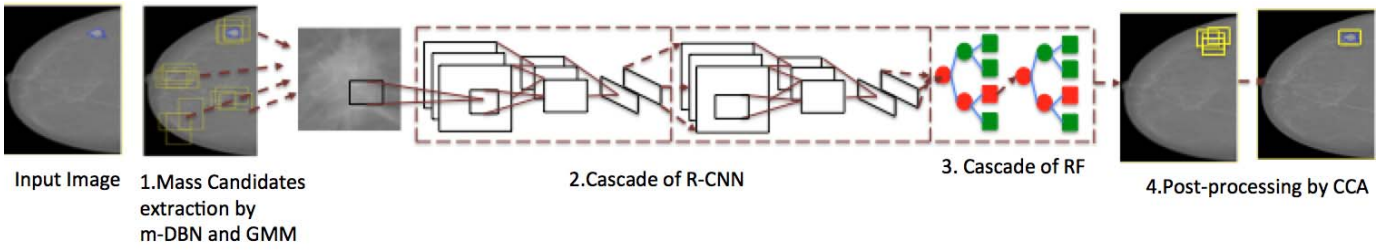

Fig. 2: Our system consists of a first stage comprising a multi-scale deep belief network (m-DBN) model and Gaussian mixture model (GMM) classifier that extract candidate regions. Subsequently, it uses cascade of R-CNN classifiers, followed by a cascade of RF classifiers and a post-processing based on CCA to detect mass regions.

literature). The third module consists of a cascade of two random forest (RF) classifiers [14] to obtain a further reduction of the number of false positives. In this stage, the input features for the first RF classifier are a set of morphological and texture features [27]. Finally a post processing step merges regions with a high overlap ratio using connected component analysis (CCA). We provide details of our proposed system below.

A. Candidate Generation with m-DBN and GMM

The m-DBN classifier [10] is trained to detect candidates in an image using a grid search over images at several different resolutions, using a mask created by a breast-air boundary threshold segmentation (using Otsu's segmentation [33]). Essentially, these m-DBN binary classifiers are trained discriminatively using input regions of fixed size across the scales, which have positive or negative labels, where the training process is based on contrastive divergence [10]. The training starts with the coarsest resolution image using all grid samples, and the samples that are classified as positive are then used to train the next (finer) resolution, forming the multi-scale cascade of DBN classifiers [10]. The inference is run at every position of the grid (i.e., every discrete position that falls within the breast mask of the respective image resolution) using the mean field approximation of the values in all DBN layers, which is followed by the computation of the free energy on the top layer [10]. In addition to m-DBN, we also use a pixel-wise GMM classification [6], [34] on the full resolution image (features are the pixel gray values), where the detection results from m-DBN and GMM are combined with CCA, using a similarity measure based on the distance between the detected pixels. The result from CCA consists of clusters of pixels being classified as belonging to a breast tissue region containing a benign and malignant mass.

B. False Positive Reduction with R-CNN

Note that the detection in Sec. III-A produces a significant number of false positives, which is usually two orders of magnitude bigger than the number of true positives. Therefore, we need a second stage that can reduce the number of false positives (see Sec. IV-A). At this second stage, we propose the use of a more complex classification methodology (compared to the first stage above) given the relatively small number of samples that remain to be processed. So we use CNN [12], [13] to extract features that are then used by a linear SVM in the region classification (this type of approach is called R-CNN [32]). A CNN [12], [13] model consists of multiple processing stages, with each stage comprising two layers (the convolutional plus activation layers, where the linear filters are applied to the image, with responses being transformed via a non-linear activation function, and the pooling and subsampling layer that reduces the input image size for the next stage - see Fig. 2), and a final stage consisting of a fully connected layer. Essentially, the convolution stages compute the output at location j from input $\mathbf{x}$ at i using the linear filter (at q^{th} stage) $\mathbf{k}^q$ and bias b^q using $\mathbf{x}(j)^q = \sigma(\sum_{i \in M_j} \mathbf{x}(i)^{q-1} * \mathbf{k}_{ij}^q + b_j^q)$, where $\sigma(.)$ is the logistic function, $*$ is the convolution operator, and M_j is the input region addresses; while the non-linear sub-sampling layers calculate subsampled data with $\mathbf{x}(j)^q = \downarrow (\mathbf{x}_j^{q-1})$, where $\downarrow (.)$ denotes a subsampling function that pools (using either the mean or max functions) the values from a region from the input data. The fully connected layer consists of the convolution equation above with a separate filter for each output location, using the whole input from the previous layer. Inference is simply the application of this process in a feed-forward manner, and training is carried out with stochastic gradient descent to minimize the classification error over the training set (via back propagation) [12], [13]. In order to compute the features from the CNN, we first crop the mass candidate with a bounding box around the candidate region from the first stage in Sec. III-A, resize the box to a fixed size of 40×40 pixels using bicubic interpolation and preprocess it with the Ball and Bruce technique [35]. Finally, we use features from the final fully connected layer of the CNN classifier and train a linear SVM. All candidates surviving the first cascade of the R-CNN are then passed through to the second cascade of R-CNN to further reduce the number of false positive as shown in the Fig.(2).

C. Final Candidate Selection by Random Forest and Post-processing

The result from the second stage presented above still contains around one order of magnitude more false positives than true positives, which need to be removed. We propose the extraction of hand-designed texture and morphological

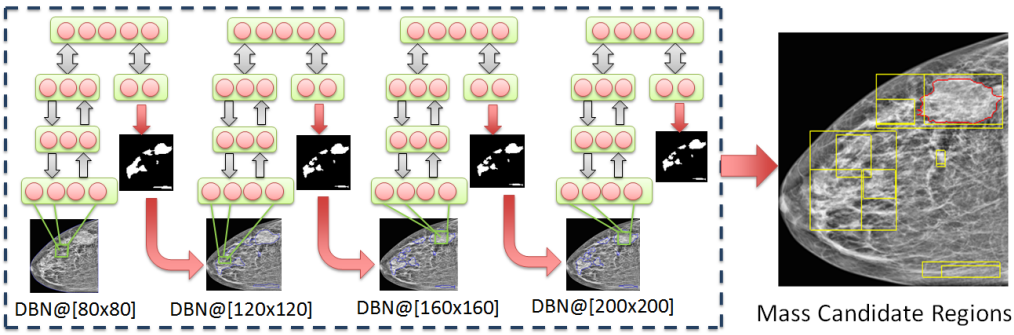

Fig. 3: Candidate generation using multi scale DBN (m-DBN).

features [27] from the remaining regions to be used in a classification process based on a random forest classifier [14]. In particular, these features include object-based measures, such as perimeter, area, perimeter-to-area ratio, circularity, rectangularity, and five normalized radial length (NRL) features [27]. The NRL features include: mean value, standard deviation, entropy and zero-crossing count [27]. Furthermore, the texture features are obtained using a gray level co-occurrence matrix (GLCM), where 13 types of features are extracted from each candidate at 14 pixel distances (from 1 to 14) and two angular directions [27]. These features are used in the training and inference processes of a two-stage cascade random forest classifier [14], where the candidates that survive the first stage are used in the training set of the second stage. Finally, the regions detected at the end of this stage are clustered using CCA using a similarity measure based on the overlap between the regions.

IV. Experiments

A. Materials and Methods

The evaluation of our methodology is carried out on two publicly available datasets: INbreast [17] and DDSM-BCRP [16]. The INbreast [17] dataset comprises a set of 115 cases containing 410 images, where 116 images contain benign or malignant masses, and the rest does not contain any masses. We compute the results using a 5-fold cross validation experiment by dividing randomly the 115 cases into 60% for training, 20% for validation and 20% for testing. All images on the DDSM-BCRP [16] dataset contain malignant masses, with 39 cases for training and 40 cases for testing. We use the free response operating characteristic curve (FROC) to calculate the number of true positive rate (TPR) as a function of false positives per image (FPI). The mass is considered to be detected if the overlap ratio between the bounding box of the candidate region and ground truth is 0.2, similar to other works in the field [20], [21], [22], [23], [24], [25], [26], [27]. The model selection for the structure of DBN, R-CNN and RF is done using the training and validation sets for INbreast, and training set only for DDSM-BCRP. We use the m-DBN as shown in Fig. 3, where the two layers contain 200 and 500 nodes and the input patch has a fixed size of 7×7 for all image scales. Recall that the m-DBN operates in a sequence of scales, which are defined as follows: the coarse resolution image is 80×80, the two intermediary scales are 120×120 and 160×160 pixels, while the finest scale has 264×264 pixels. For the R-CNN cascade classifiers, we use the LeNet network structure [12]. We artificially augment the number of positive samples by translating (64 shifts) and rotating (three angular rotation namely 90, 180, 270 degrees) the bounding box from positive candidates. The new augmented training set contains samples which are approximately 10 times the training samples in original training set. The desired operating point during the training of each module in the system is fixed to be $TPR \geq 0.90$, while gradually reducing the number of false positives per image until the final stage of post-processing where there is around 1 false positive per image. The run time efficiency is is assessed with the test time for detection using a standard computer (Intel(R) Core(TM) i5-2500k 3.30GHz CPU with 8GB RAM).

On average, using the results on the test sets obtained on INbreast and DDSM-BCRP, our method generates 300 mass candidates from the first stage (m-DBN + GMM, followed by CCA clustering). During this first stage, our method has a TPR of one, which means that we detect all the masses. The result from the first stage of the R-CNN cascade is shown in the 4(a). After the first stage, the number of false positive candidates per image is reduced to around 50 which subsequently gets reduced to 30 in the second stage of R-CNN as shown in 4(b). The results of the first and second stages of the RF cascade are shown in Fig. 4(c) and 4(d), respectively. It shows that the first stage of RF will reduce the false positives per image to around eight and after the final stage, the number of false positives per image is reduced to around five. The TPR in the FROC curve shown in 4(a-d) is computed such that only one detected candidate with the highest overlap ratio with the ground-truth is considered as a true positive (with an overlap ratio bigger than 0.2, as before).

The FROC with the error bar plot, shown in Fig. 5(a), describes the performance of our system using the five-fold cross validation experiment on the INbreast data using the whole cascade after the post-processing using CCA. In general on INbreast, our true positive performance saturates on the test

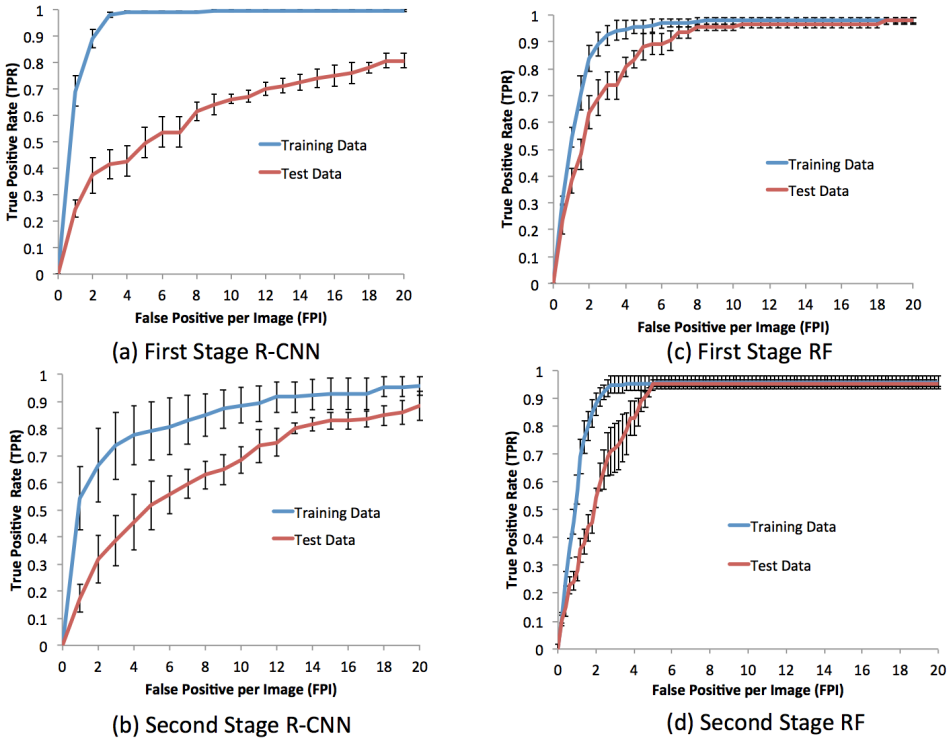

Fig. 4: FROC curve showing the result on various operating points with true positive rate (TPR) versus false positive per image (FPI) on INbreast dataset after each cascade ((a) and (b) is the result from the first module containing cascade of R-CNN whereas (c) and (d) shows the result of random forest on morphological and texture features) of our system.

set at TPR of 0.96 ± 0.03 at FPI $= 1.2$ and TPR of 0.94 ± 0.02 at FPI $= 0.3$ for the training data. The FROC curve on DDSM-BCRP shows only the average result (i.e, no error bars with standard deviation) on the single public train and test sets in Fig. 5(b). Tab. I shows a performance comparison of several state-of-the-art methods for mass detection in mammograms, where the results from the competing methods are as reported by Horsh et al. [28] or by their original authors. However, note that the majority of the results on DDSM dataset cannot be directly compared to ours because they have been obtained with an experimental setup that is not publicly available, and so cannot be reproduced (indicated by the column "Rep."). Also in Tab. I, the acronym "NA" [28] indicates that the respective result is unavailable, and not all competing methodologies are tested in mammograms containing all possible types of masses (benign, normal and malignant - indicated by "all") - instead they are tested using only a subset of the images containing malignant masses (indicated by "malig"). Finally, we show some example results produced by our system on Fig. 6 on images displaying varying degrees of detection difficulty.

V. Discussion and Conclusions

From the results shown in Fig. 5 and Tab.(I), we notice that our method produces the best results in the field (by a large margin) for the INbreast and DDSM-BCRP datasets. One of the important observations made during the training of our system is that in order to get state-of-the-art results on INbreast, the FPI from the second stage should be kept under 30 per image. We also observed that a single R-CNN (i.e., a single-stage cascade) is not able to reduce the FPI from 300 to 30, but the combination of two cascades of R-CNN achieves this goal, while keeping the TPR above 0.9. Similarly, we also notice that a single-stage cascade of RF is not able to reduce the FPI to around five in both datasets, but the addition of a second stage of RF reaches that objective. The CCA labelling significantly removes the false positive per image from five to around one, keeping the TPR ≥ 0.9, as the detection from the second cascade of RF contains around three-four regions per image that are overlapping with TPR.

The comparison with competing methods in Table I shows that our methodology currently produces the best results to date in both datasets: TPR $= 0.96 \pm 0.03$ @ 1.2 FPI and TPR = 0.87 ± 0.14 @ 0.8 FPI for INbreast; and TPR $= 0.75$ @ 4.8 FPI and TPR $= 0.70$ @ 4 FPI for DDSM-BCRP. Moreover, our methodology compares well to the competing methods with respect to the inference time, where our method runs in 20 seconds, compared to the 108 seconds needed for one of the competing methods [20] on INbreast. There are some important notes to make about the training process that are not displayed in the results: 1) different types of CNN structures including different filter sizes have been tried but did not show significant improvement. We also tried to add more than two

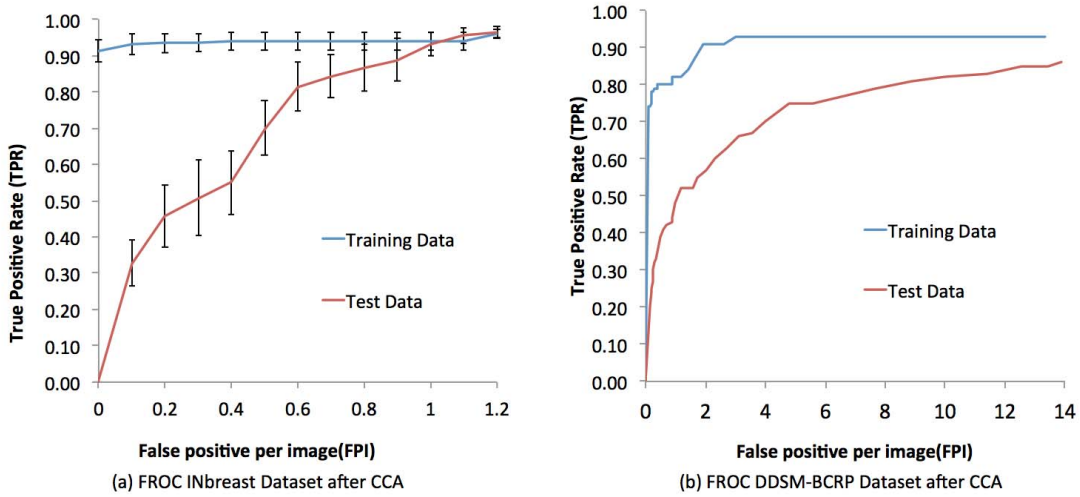

Fig. 5: FROC curve showing the result on various operating points with true positive rate (TPR) versus false positive per image (FPI) on INbreast (a) and DDSM-BCRP (b) after CCA post processing from the final cascade.

TABLE I: Comparison between different state-of-the-art methodologies.

Method	Images	Rep.	Dataset	TPR@FPI	Type	Running time
Our method	410	yes	INbreast	0.96±0.03@1.2,0.87±0.14@0.8	all	20s
Kozegar et al. [20]	116	yes	INbreast	0.87@3.67	malig	108 s
Our method	316	yes	DDSM-BCRP	0.75@4.8, 0.70@4	all	20s
Beller et al. [21]	160	yes	DDSM-BCRP	0.70@8	malig	NA
Campanini et al. [23]	512	no	DDSM	0.80@1.1	malig	NA
Eltonsy et al. [24]	270	no	DDSM	0.92@5.4, 0.88@2.4, 0.81@0.6	all	NA
Sampat et al. [25]	100	no	DDSM	0.88@2.7, 0.85@1.5,0.8@1	all	NA
Wei et al. [27]	400	no	University of Michigan	0.70@0.79,0.8@1.2,0.9@2	all	NA
Petrick et al. [36]	25	no	University of Michigan	0.96@4.5,0.96@3.0	all	NA
Mudigonda et al. [37]	56	yes	MIAS	0.81@2.3	all	NA
Brake et al. [22]	772	no	Nijmegen	0.70@0.10	malig	NA
Bellotti et al. [26]	3369	no	MAGIC-5	0.80@4.23	all	NA
Yin et al. [19]	92	no	University of Chicago	0.95@3	all	NA

stages of cascade in the second stage of our method (R-CNN), which also did not show significant improvement; 2) for the m-DBN model, we have also tried different input sizes: 3×3 ,$5 \times 5$and 7×7 patches, but the latter produced the best results.

Finally, from the visual results in Fig(6), we can see that our system produces a visually accurate detection result in most test images. The images in Fig. 6(a-n) contain a single mass and our system is able to detect all of them without any false positives. The image in Fig 6(o-u) represents a normal mammogram without any masses and our system does not find any FP, which suggests that our system is robust to normal mammograms without findings. The result in Fig. 6(v and w) contains a large mass and smaller mass, which was detected by our system along with two false positives. Finally, we show an false positive detection in Fig. 6(x and y), where our system detects a possible mass in a normal image. The main issue currently affecting our method is the limited size of the training set at the later stages, where not many samples remain to train our model. In future we would like to have fully automated detection, segmentation and classification of the mass using system based on what we have in this paper.

REFERENCES

[1] A. Jemal, R. Siegel, E. Ward, Y. Hao, J. Xu, T. Murray, and M. J. Thun, "Cancer statistics, 2008," *CA: a cancer journal for clinicians*, vol. 58, no. 2, pp. 71–96, 2008.

[2] A. I. of Health and W. . C. Australia, "Breast cancer in australia: an overview," 2012.

[3] A. Oliver, J. Freixenet, J. Marti, E. Perez, J. Pont, E. R. Denton, and R. Zwiggelaar, "A review of automatic mass detection and segmentation in mammographic images," *Medical Image Analysis*, vol. 14, no. 2, pp. 87–110, 2010.

[4] J. Tang, R. M. Rangayyan, J. Xu, I. El Naqa, and Y. Yang, "Computer-aided detection and diagnosis of breast cancer with mammography: recent advances," *Information Technology in Biomedicine, IEEE Transactions on*, vol. 13, no. 2, pp. 236–251, 2009.

[5] C. Dromain, B. Boyer, R. Ferre, S. Canale, S. Delaloge, and C. Balleyguier, "Computed-aided diagnosis (cad) in the detection of breast cancer," *European journal of radiology*, vol. 82, no. 3, pp. 417–423, 2013.

[6] N. Dhungel, G. Carneiro, and A. P. Bradley, "Deep structured learning for mass segmentation from mammograms," *arXiv preprint arXiv:1410.7454*, 2014.

[7] M. L. Giger and A. Pritzker, "Medical imaging and computers in the diagnosis of breast cancer," in *SPIE Optical Engineering+ Applications*. International Society for Optics and Photonics, 2014, pp. 918 908–918 908.

[8] J. G. Elmore, S. L. Jackson, L. Abraham *et al.*, "Variability in interpretive performance at screening mammography and radiologists

characteristics associated with accuracy1," *Radiology*, vol. 253, no. 3, pp. 641–651, 2009.

[9] C. Balleyguier, K. Kinkel, J. Fermanian, S. Malan, G. Djen, P. Taourel, and O. Helenon, "Computer-aided detection (cad) in mammography: does it help the junior or the senior radiologist?" *European journal of radiology*, vol. 54, no. 1, pp. 90–96, 2005.

[10] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[11] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 1–38, 1977.

[12] Y. LeCun and Y. Bengio, "Convolutional networks for images, speech, and time series," *The handbook of brain theory and neural networks*, vol. 3361, 1995.

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks." in *NIPS*, vol. 1, no. 2, 2012, p. 4.

[14] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[15] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim, "Do we need hundreds of classifiers to solve real world classification problems?" *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 3133–3181, 2014.

[16] M. Heath, K. Bowyer, D. Kopans, R. Moore, and P. Kegelmeyer, "The digital database for screening mammography," in *Proceedings of the 5th international workshop on digital mammography*, 2000, pp. 212–218.

[17] I. C. Moreira, I. Amaral, I. Domingues, A. Cardoso, M. J. Cardoso, and J. S. Cardoso, "Inbreast: toward a full-field digital mammographic database," *Academic Radiology*, vol. 19, no. 2, pp. 236–248, 2012.

[18] J. J. Fenton, S. H. Taplin, P. A. Carney *et al.*, "Influence of computer-aided detection on performance of screening mammography," *New England Journal of Medicine*, vol. 356, no. 14, pp. 1399–1409, 2007.

[19] F.-F. Yin, M. L. Giger, K. Doi, C. J. Vyborny, and R. A. Schmidt, "Computerized detection of masses in digital mammograms: Automated alignment of breast images and its effect on bilateral-subtraction technique," *Medical Physics*, vol. 21, no. 3, pp. 445–452, 1994.

[20] E. Kozegar, M. Soryani, B. Minaei, I. Domingues *et al.*, "Assessment of a novel mass detection algorithm in mammograms," *Journal of cancer research and therapeutics*, vol. 9, no. 4, p. 592, 2013.

[21] M. Beller, R. Stotzka, T. O. Müller, and H. Gemmeke, "An example-based system to support the segmentation of stellate lesions," in *Bildverarbeitung für die Medizin 2005*. Springer, 2005, pp. 475–479.

[22] G. M. te Brake, N. Karssemeijer, and J. H. Hendriks, "An automatic method to discriminate malignant masses from normal tissue in digital mammograms," *Physics in Medicine and Biology*, vol. 45, no. 10, p. 2843, 2000.

[23] R. Campanini, D. Dongiovanni, E. Iampieri, N. Lanconelli, M. Masotti, G. Palermo, A. Riccardi, and M. Roffilli, "A novel featureless approach to mass detection in digital mammograms based on support vector machines," *Physics in Medicine and Biology*, vol. 49, no. 6, p. 961, 2004.

[24] N. H. Eltonsy, G. D. Tourassi, and A. S. Elmaghraby, "A concentric morphology model for the detection of masses in mammography," *Medical Imaging, IEEE Transactions on*, vol. 26, no. 6, pp. 880–889, 2007.

[25] M. P. Sampat, A. C. Bovik, G. J. Whitman, and M. K. Markey, "A model-based framework for the detection of spiculated masses on mammographya)," *Medical physics*, vol. 35, no. 5, pp. 2110–2123, 2008.

[26] R. Bellotti, F. De Carlo, S. Tangaro, G. Gargano, G. Maggipinto, M. Castellano, R. Massafra, D. Cascio, F. Fauci, R. Magro *et al.*, "A completely automated cad system for mass detection in a large mammographic database," *Medical physics*, vol. 33, no. 8, pp. 3066–3075, 2006.

[27] J. Wei, B. Sahiner, L. M. Hadjiiski, H.-P. Chan, N. Petrick, M. A. Helvie, M. A. Roubidoux, J. Ge, and C. Zhou, "Computer-aided detection of breast masses on full field digital mammograms," *Medical physics*, vol. 32, no. 9, pp. 2827–2838, 2005.

[28] A. Horsch, A. Hapfelmeier, and M. Elter, "Needs assessment for next generation computer-aided mammography reference image databases and evaluation studies," *International journal of computer assisted radiology and surgery*, vol. 6, no. 6, pp. 749–767, 2011.

[29] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[30] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," *arXiv preprint arXiv:1409.4842*, 2014.

[31] H. Li, Z. Lin, X. Shen, J. Brandt, and G. Hua, "A convolutional neural network cascade for face detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 5325–5334.

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

[33] N. Otsu, "A threshold selection method from gray-level histograms," *Automatica*, vol. 11, no. 285-296, pp. 23–27, 1975.

[34] N. Dhungel, G. Carneiro, and A. P. Bradley, "Tree re-weighted belief propagation using deep learning potentials for mass segmentation from mammograms."

[35] J. E. Ball and L. M. Bruce, "Digital mammographic computer aided diagnosis (cad) using adaptive level set segmentation," in *Engineering in Medicine and Biology Society, 2007. EMBS 2007. 29th Annual International Conference of the IEEE*. IEEE, 2007, pp. 4973–4978.

[36] N. Petrick, H.-P. Chan, B. Sahiner, and D. Wei, "An adaptive density-weighted contrast enhancement filter for mammographic breast mass detection," *Medical Imaging, IEEE Transactions on*, vol. 15, no. 1, pp. 59–67, 1996.

[37] N. R. Mudigonda, R. M. Rangayyan, and J. L. Desautels, "Detection of breast masses in mammograms by density slicing and texture flow-field analysis," *Medical Imaging, IEEE Transactions on*, vol. 20, no. 12, pp. 1215–1227, 2001.

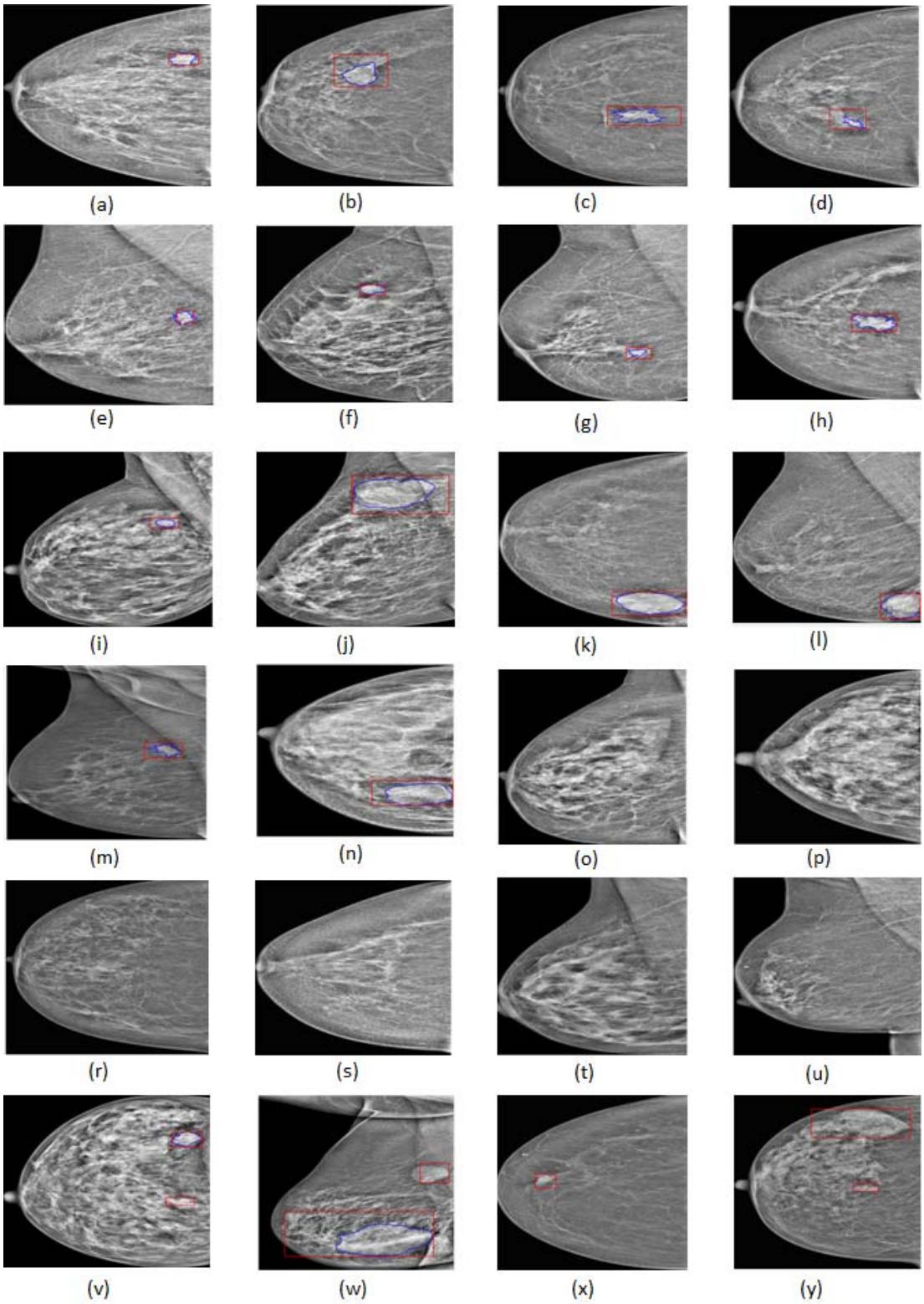

Fig. 6: Some few examples of our mass detection system from test data in INbreast dataset. Red box is the bounding box generated by our detection algorithm whereas blue lines denotes the contour of the ground truth. Note that all the mammograms are segmented to breast air boundary using Otsu [33] method and are aligned in one direction.