Automated Mass Detection from Mammograms using Deep Learning and Random Forest

Neeraj Dhungel[1] Gustavo Carneiro[1] Andrew P. Bradley[2]

[1] ACVT, University of Adelaide, Australia
[2] University of Queensland, Australia *

Abstract. Mass detection from mammogram plays an crucial role as a pre-processing stage for mass segmentation and classification. In this paper, we present a novel approach for detecting masses from mammograms using a cascade of deep learning and random forest classifiers. The deep learning classifier consists of a multi-scale deep belief network classifier that selects regions to be further processed by a two-level cascade of deep convolutional neural networks. The regions that survive this deep learning analysis is then processed by a two-level cascade of random forest classifiers that use several morphological and texture features extracted from those surviving regions. We show that the proposed cascade of deep learning and random forest classifiers are effective in the reduction of false positive regions, while keeping a high true positive detection, and that the final mass detection produced by our approach achieves the best results in the field on public mammogram datasets.

Keywords: mass detection, multi scale processing, deep learning, morphological and texture features, random forest

1 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 death worldwide. Mammography is a widely used imaging modality for screening breast cancer because it enables the detection of suspicious lesions (e.g., masses), which is the first step in assessing the risk of having, or developing breast cancer. However, the problem is that, nowadays, this is mostly a manual process, where a significant number of breast masses are missed or those which are detected turns out to be benign after their biopsies [2]. In part, this happens because of the masses' variability in shape, size and boundary [3, 2], and also because of their low signal-to-noise ratio 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. An automated mass detection system is useful in clinical practice to provide a "second" opinion that can help improve the mammogram analysis consistency, and as a result, reduce to some extent, the current dependence on the radiologist's experience and workload [4].

* 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).

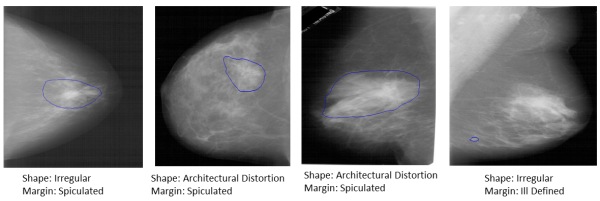


Fig. 1: Some categories of malignant masses from DDSM-BCRP dataset.

In spite of the development of numerous mass detection systems, they are still not widely used in clinical practice because they tend to reduce the accuracy of the radiologists [5]. Most of the 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 [6–13]. This first stage is followed by a false positive removal step, using different types of classifiers, such as support vector machine, linear discriminant analysis or neural network [6–13]. 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 [2]. Another problem is that they are usually tested on private datasets or on random subsets of DDSM [14], which makes direct comparisons difficult [15]. In addition to this, most of the current methods [6–9] are tested only in mammograms that contain malignant masses, which creates a bias in the results.

Our 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 forest (see Fig. 2). In our method, the first stage consists of a multi-scale deep belief network (m-DBN) cascade of classifiers [16] combined with a Gaussian mixture model [17](GMM) classifier that select a set of regions, representing candidates for containing breast masses. This first step is followed by a second stage, comprising a cascade of deep convolutional neural networks [18, 19] that reduces the number of false postives, while keeping the large majority of the true positive regions. Finally, we extract texture and morphological features from the remaining regions to be classified by a random forest classifier [20]. We show that our methodology produces the best results in the field on the public datasets DDSM-BCRP [14] and INbreast [21] using mammograms containing no findings, and malignant and benign masses.

2 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) [16, 3], shown in Fig. 3, with a Gausssian mixture model [17](GMM) classifier for candidate generation. These candidates are the inputs to a second step containing a cascade of two stages of deep convolutional neural networks (CNN) [18, 19] that produce features for being used by a linear support vector machine (SVM) classifier (this combination of CNN and SVM is known as an R-CNN [22] in computer vision literature). The third module consists of a cascade of two stages of random forest (RF) [20] for further reduction of false positives, where the input features for the RF classifier is a set of texture and morphological fearures. Finally the fourth

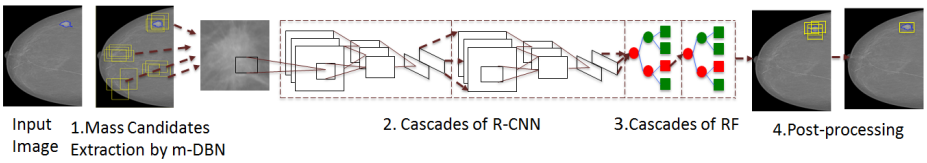


Fig. 2: Our system consists of a first stage comprising an m-DBN and GMM that extract candidate regions. Subsequently, it uses 2 cascades of R-CNN, followed by an RF classifier and a post-processing based on CCA to detect mass regions.

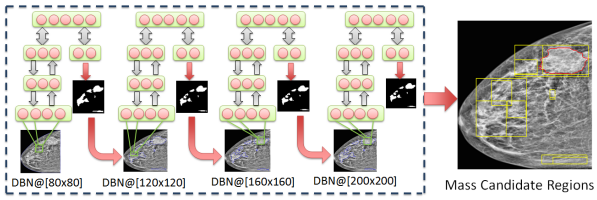


Fig. 3: Candidate generation using multi scale DBN (m-DBN).

module is a post processing module that merges regions with high overlap ratio using connected component analysis (CCA). We provide details of our system in this section.

2.1 Candidate Generation with m-DBN and GMM

The m-DBN classifier [16] is trained to detect candidates in an image using a grid search over images at four different resolutions, using a mask created by a breast-air boundary segmentation (using Otsu's segmentation [23]). Specifically, assuming that the full resolution image has 264×264 pixels, the other three resolutions are: 160×160, 120×120 and 80×80. Essentially, these m-DBN binary classifiers are trained discriminatively using input regions of size 7×7 (fixed across the scales), which have positive or negative labels, where the training process is based on contrastive divergence [16]. This 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 [16]. The inference is run in 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 [16]. In addition to m-DBN, we also use a pixel-wise GMM classification [3] 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 region containing a mass.

2.2 False Positive Reduction with R-CNN

Note that the detection in Sec. 2.1 still produces a significant amount of false positives, which is usually two orders of magnitude bigger than the number of true positives, so we need a second stage that can reduce this amount of false positives. 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 [18, 19] to extract features that are then used by a linear SVM in the region classification (this type of approach is called R-CNN [22] and has produced state-of-the-art results in object detection and semantic segmentation). A CNN [18, 19] 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) [18, 19]. 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. 2.1, resize the box to a fixed size of 40×40 pixels using bicubic interpolation and preprocess it with the Ball and Bruce technique [24]. 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 false positive as shown in the Fig.(2).

2.3 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 features [13] from the remaining regions to be used in a classification process based on random forest [20]. In particular, these features include object-based measures, such as number of perimeter pixels, area, perimeter-to-area ratio, circularity, rectangularity, and five normalized radial length (NRL) features [13]. The NRL features include: mean value, standard deviation, entropy, area ratio and zero-crossing count [13]. Furthermore, the texture features are obtained using gray level co-occurrence matrix (GLCM), where thirteen types of features are extracted from each candidate at fourteen pixel distances and two angular directions [13]. These features are used in the training and inference processes of a two-stage cascade random forest classifier [20], where the candidates that

survive the first stage are used in the training 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.

3 Experiments

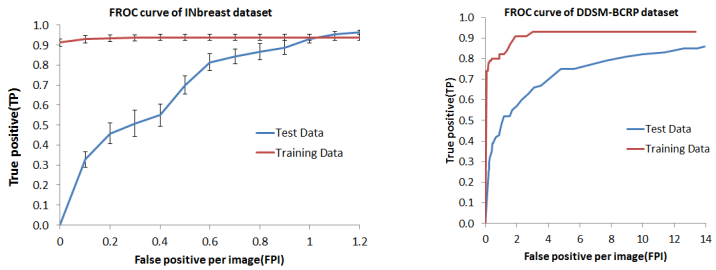
3.1 Materials and Methods

The evaluation of our methodology is carried out on two publicly available datasets: INbreast [21] and DDSM-BCRP [14]. The INbreast [21] dataset comprises a set of 115 cases containing 410 images, where out of 410 images, 116 images contains the benign or malignant masses, whereas the rest does not contain any masses. We run a five-fold cross validation experiment on INbreast, where we divide the images in terms of the 115 cases in a mutually exclusive manner, with 60% of the cases for training, 20% for validation and 20% for testing. All images on the DDSM-BCRP [14] dataset contain malignant masses, with 39 cases for training and 40 cases for testing. We use the free receiver operating curve (FROC) to calculate number of true positive (TP) at given false positive per image(FPI). Efficiency is calculated with test time detection using a standard computer (Intel(R) Core(TM) i5-2500k 3.30GHz CPU with 8GB RAM). The mass is considered to be detected if the overlap ratio between the bounding box of candidate region and ground truth is 0.2, similar to other works in the field [6–13]. The model selection for the structure of DBN, R-CNN and RF is done via cross validation using training and validation sets for INbreast, and training 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 size 7×7 for all image scales. For the R-CNN cascade classifiers, we use the LeNet network structure [18]. We artificially augment the number of positive samples by translating and rotating the bounding box from positive candidates. The operating point during the training of each module in the system is fixed to be $TP \geq 0.90$, while gradually reducing the number of false positives per image until the last module.

3.2 Results

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 TP rate of 1, which means that we never miss any of the masses. After the second stage, the number of candidates is reduced to around 20, and after the final stage, the number of false positives per image is reduced to around 2.

The FROC with the error bar plot (indicating mean and standard deviation results), shown in Fig. 4(a), describes the performance of our system using the five-fold cross validation experiment on INbreast. In general on INbreast, our true positive performance saturates on the test set at TP of 0.96 ± 0.03 at FPI $= 1.2$ and TP of 0.94 ± 0.02 at FPI $= 0.3$ for the training data. The FROC on DDSM-BCRP shows only the average result on the suggested train and test sets in Fig. 4(b). Tab. 1 shows a performance comparison of several state-of-the-art methods for mass detection in mammograms, where the results from the other methods are as reported by Horsh et al. [15] or by their original authors. However, note that the majority of the results on DDSM dataset cannot be compared directly to ours because they have been obtained with an experimental setup that



(a) FROC curve of INbreast (b) FROC curve of DDSM-BCRP

Fig. 4: FROC curve showing the result on various operating point with true positive(TP) against false positive per image(FPI) on INbreast (a) and DDSM-BCRP (b).

Table 1: Comparision between different state-of-the-art methodologies.

Method	Images	Rep.	Dataset	TP@FPI	Type	Time
Our method	410	yes	INbreast	0.96±0.03@1.2,0.87±0.14@0.8	all	20s
Kozegar et al. [6]	116	yes	INbreast	0.87@3.67	malig	108 s
Our method	158	yes	DDSM-BCRP	0.75@4.8, 0.70@4	all	20s
Beller et al. [7]	160	yes	DDSM-BCRP	0.70@8	malig	NA
Campanini et al. [9]	512	no	DDSM	0.80@1.1	malig	NA
Eltonsy et al. [10]	270	no	DDSM	0.92@5.4, 0.88@2.4, 0.81@0.6	all	NA
Sampat et al. [11]	100	no	DDSM	0.88@2.7, 0.85@1.5,0.8@1	all	NA
Brake et al. [8]	772	no	Nijmegen	0.70@0.10	malig	NA
Bellotti et al. [12]	3369	no	MAGIC-5	0.80@4.23	all	NA
Wei et al. [13]	400	no	Uni. of Michigan	0.70@0.79,0.8@1.2,0.9@2	all	NA

is not publicly available, and so cannot be reproduced (indicated by the column "Rep."). Also in Tab. 1, the acronym "NA" [15] indicates performance measures that are unavailable, and not all other 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 illustrative example of results produced by our system in Fig. 5.

4 Discussion and Conclusions

From the results shown in Fig. 4 and Tab.(1), 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 our system is that in order to get the state-of-the-art results on INbreast, the FPI from the second stage should be kept under 20 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 20, but the combination of two cascades of R-CNN achieves this goal, while keeping the TP above 0.9. Similarly, we also notice that a single-stage cascade of RF is not able to reduce the FPI to around 2-3 in both datasets, but the addition of a second stage of RF reaches that objective. The

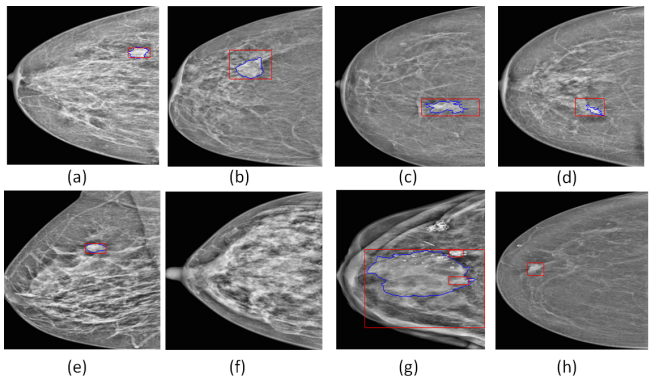


Fig. 5: Some few examples of our mass detection system from test data in INbreast dataset. Red box is the bounding box generated by our detection algorithm whereas blue lines denotes the contour of the ground truth.

comparison with other methods in Table 1 shows that our methodology currently produces the best results in both datasets: TP $= 0.96 \pm 0.03$ @ 1.2 FPI and TP $= 0.87 \pm 0.14$ @ 0.8 FPI for INbreast; and TP $= 0.75$ @ 4.8 FPI and TP $= 0.70$ @ 4 FPI for DDSM-BCRP. Moreover, our methodology compares well to others with respect to the inference time, where we have 20 seconds against 108 seconds for a competing method [6] 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. We also tried to add more than two stages of cascade in the second stage of our method (R-CNN), but it did not show significant improvement; 2) for the m-DBN model, we have also tried different input sizes: 3×3 and 7×7 patches, but the latter produced the best results. Finally, from the visual results in Fig(5), we can see that our system produces an accurate detection result in test images at the operating point of TP $= 0.96 \pm 0.03$TP@1.2FPI. The images in Fig. 5(a-e) contain a single mass and our system is able to detect all of them without any false positve. The image in Fig 5(f) represents a normal mammogram without any mass and our system does not find any FP, which suggests that our system is robust to mammograms without findings. The result in Fig. 5(g) contains a large mass, which was detected by our system along with two false positives (internal to the bounding box of this large mass) - note that we cannot remove these smaller masses using the CCA post-processing because the overlap must be large for both masses (not only one of them, as shown in Fig. 5(g)). Finally, we show an FP detection in Fig. 5(h) where our system detects a false positive mass in a normal image. The main issue currently affecting our method is the limited size of the training sets at the later stages, where not many samples remain to train our model.

References

1. Jemal, A. et al.: Cancer statistics, 2008. CA: a cancer journal for clinicians **58**(2) (2008) 71–96

2. Oliver, A., et al.: A review of automatic mass detection and segmentation in mammographic images. MedIA **14**(2) (2010) 87–110
3. Dhungel, N., Carneiro, G., Bradley, A.P.: Deep structured learning for mass segmentation from mammograms. arXiv preprint arXiv:1410.7454 (2014)
4. Elmore, J.G., Jackson, S.L., Abraham, L., et al.: Variability in interpretive performance at screening mammography and radiologists characteristics associated with accuracy1. Radiology **253**(3) (2009) 641–651
5. Fenton, J.J., Taplin, S.H., Carney, P.A., et al.: Influence of computer-aided detection on performance of screening mammography. New England Journal of Medicine **356**(14) (2007) 1399–1409
6. Kozegar, E., Soryani, M., Minaei, B., Domingues, I., et al.: Assessment of a novel mass detection algorithm in mammograms. Journal of cancer research and therapeutics **9**(4) (2013) 592
7. Beller, M., Stotzka, R., Müller, T.O., Gemmeke, H.: An example-based system to support the segmentation of stellate lesions. In: Bildverarbeitung für die Medizin 2005. Springer (2005) 475–479
8. te Brake, G.M., Karssemeijer, N., Hendriks, J.H.: An automatic method to discriminate malignant masses from normal tissue in digital mammograms. Physics in Medicine and Biology **45**(10) (2000) 2843
9. Campanini, R., et al.: A novel featureless approach to mass detection in digital mammograms based on support vector machines. Physics in Medicine and Biology **49**(6) (2004) 961
10. Eltonsy, N.H., Tourassi, G.D., Elmaghraby, A.S.: A concentric morphology model for the detection of masses in mammography. Medical Imaging, IEEE Transactions on **26**(6) (2007) 880–889
11. Sampat, M.P., Bovik, A.C., Whitman, G.J., Markey, M.K.: A model-based framework for the detection of spiculated masses on mammographya). Medical physics **35**(5) (2008) 2110–2123
12. Bellotti, R., De Carlo, F., Tangaro, S., Gargano, G., Maggipinto, G., Castellano, M., Massafra, R., Cascio, D., Fauci, F., Magro, R., et al.: A completely automated cad system for mass detection in a large mammographic database. Medical physics **33**(8) (2006) 3066–3075
13. Wei, J., et al.: Computer-aided detection of breast masses on full field digital mammograms. Medical physics **32**(9) (2005) 2827–2838
14. Heath, M., Bowyer, K., Kopans, D., Moore, R., Kegelmeyer, P.: The digital database for screening mammography. In: Proceedings of the 5th international workshop on digital mammography. (2000) 212–218
15. Horsch, A. et al.: Needs assessment for next generation computer-aided mammography reference image databases and evaluation studies. International journal of computer assisted radiology and surgery **6**(6) (2011) 749–767
16. Hinton, G.E., Salakhutdinov, R.R.: Reducing the dimensionality of data with neural networks. Science **313**(5786) (2006) 504–507
17. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society. Series B (Methodological) (1977) 1–38
18. LeCun, Y., Bengio, Y.: Convolutional networks for images, speech, and time series. The handbook of brain theory and neural networks **3361** (1995)
19. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS. Volume 1. (2012) 4
20. Breiman, L.: Random forests. Machine learning **45**(1) (2001) 5–32
21. Moreira, I.C., et al.: Inbreast: toward a full-field digital mammographic database. Academic Radiology **19**(2) (2012) 236–248
22. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: CVPR 2014
23. Otsu, N.: A threshold selection method from gray-level histograms. Automatica **11**(285-296) (1975) 23–27
24. Ball, J.E., Bruce, L.M.: Digital mammographic computer aided diagnosis (cad) using adaptive level set segmentation. In: EMBS 2007