Eurographics Workshop on Visual Computing for Biology and Medicine (2017)
S. Bruckner, A. Hennemuth, and B. Kainz (Editors)

Mammogram Classification and Abnormality Detection from Nonlocal Labels using Deep Multiple Instance Neural Network

Yoni Choukroun[1†], Ran Bakalo[2], Rami Ben-Ari[2], Ayelet Askelrod-Ballin[2], Ella Barkan[2] and Pavel Kisilev[2‡]

[1]Computer Science Department, Technion - Israel Institute of Technology
[2]IBM Research - Haifa

Abstract

Mammography is the common modality used for screening and early detection of breast cancer. The emergence of machine learning, particularly deep learning methods, aims to assist radiologists to reach higher sensitivity and specificity. Yet, typical supervised machine learning methods demand the radiological images to have findings annotated within the image. This is a tedious task, which is often out of reach due to the high cost and unavailability of expert radiologists. We describe a computer-aided detection and diagnosis system for weakly supervised learning, where the mammogram (MG) images are tagged only on a global level, without local annotations. Our work addresses the problem of MG classification and detection of abnormal findings through a novel deep learning framework built on the multiple instance learning (MIL) paradigm. Our proposed method processes the MG image utilizing the full resolution, with a deep MIL convolutional neural network. This approach allows us to classify the whole MG according to a severity score and localize the source of abnormality in full resolution, while trained on a weakly labeled data set. The key hallmark of our approach is automatic discovery of the discriminating patches in the mammograms using MIL. We validate the proposed method on two mammogram data sets, a large multi-center MG cohort and the publicly available INbreast, in two different scenarios. We present promising results in classification and detection, comparable to a recent supervised method that was trained on fully annotated data set. As the volume and complexity of data in healthcare continues to increase, such an approach may have a profound impact on patient care in many applications.

1. Introduction

The most common cancer and second leading cause of death among women is breast cancer [JBC*11] where the medical community is striving for its early detection. Mammography is commonly used for screening and detection of breast cancer [CC05]. In current practice, radiologists and CADx systems follow a two-stage process defined by the detection of abnormalities, followed by their classification according to the standard Breast Imaging Reporting and Data System (BI-RADS). Mammogram analysis is challenging due to the high variability of breast patterns, variations in appearance, size and shape of the abnormalities, which often make them difficult to detect and classify even by expert radiologists. Subsequently, a broad variety of traditional machine learning classifiers have been developed for automatic diagnosis of different findings such as masses and calcifications, and ultimately breast cancer [dOdCFS*15, JY15]. Traditional computer vision and machine learning approaches suggest task-specific and handcrafted features for classification and detection problems. Deep learning algorithms outperform these methods in many areas and have solved complicated pattern recognition problems, especially in the domain of Big Data [SZ14]. The deep architecture can be used to efficiently learn optimal representations, and ultimately to enhance detection and classification accuracy. Deep learning methods have also achieved considerable success in multiple applications, including medical imaging, even where the available data is limited [YCBL14, BAABKH17].

Several studies [CGGS13, PND*14, KPN*16] have explored deep learning approaches to address the automatic classification of lesions in mammography. However, one hurdle to fully utilizing the potential of Big Data in medical images is the expensive process of annotating images, which poses a strong bottleneck in supervised learning and particularly in medical imaging. In the weakly supervised paradigm, only image-level tags are necessary to train a classifier, as opposed to fully supervised classification and detection, which typically requires exhaustive annotations of the medical images [DCB16, DCB17]. Yet, there are several other reasons justifying the weakly labeled approach in medical imaging:

- The tedious annotation produces an additional source for errors in data labels
- Often lesion margins are ambiguous, creating controversial annotations

† Research conducted as an intern at IBM Research - Haifa.
‡ Now at Huawei Research - Israel.

- Global labels for training are commonly available through medical records and easy to acquire
- Data-driven class discrimination can lead to new insights
- Weakly methods can be used to complement fully supervised approaches

Therefore, medical image classification and localization of abnormalities with only global tags, is a highly desired task.

Weakly supervised models have recently gained high interest in the computer vision community [BV16, TWH*17] and in medical imaging [HK16, YZP*16, QLC*16], facilitating technological advancements in computer-aided diagnosis. We address the task of mammogram (MG) classification and abnormality detection, as illustrated in Fig. 1. The associated lesions with each class determine the classification task as well as the type of alerted findings.

Several studies address the problem of weakly supervised learning in MG *e.g.* [QLC*16, HK16, ZLVX16]. Quellec *et al.* [QLC*16] use hand crafted features to specifically capture masses and micro-calcifications and test their method on the digitally **scanned** image data set of Digital Database for Screening Mammography (DDSM) [OGA*08] unlike the currently used Full Field Digital Mammogrpahy (FFDM). Recent studies in medical imaging for weakly supervised classification tasks, include [HK16, YZP*16, ZLVX16]. Yan *et al.* [YZP*16] address a totally different problem of anatomy recognition in CT scans using MIL. This work uses a cascade classifier and focuses on background classification (referred as non-informative patches). The patches are extracted from CT slices and used to boost learning for recognition of the large body parts, appearing in the slice images. Hwang and Kim [HK16] address MG classification on weakly labeled sets using a CNN, divided into two branches, one for classification and the other for detection. Their network works on downsized images (500 $\times$ 500) with low resolution in localization, and strongly impacting the low classification performance reaching 0.675 ROC-Area Under the Curve, with results reported on digitally scanned images.

Multiple Instance Learning (MIL) suggests an approach for weakly supervised learning, namely when the labels are provided at the whole image level. By representing an image as a *bag* of multiple *instances*, classification can be made based on bag ingredient labels instead of the traditional global image features [DLLP97, MLP98]. MIL has been regained interest recently in the deep learning framework for weakly supervised object recognition [WYHY15, SHLKR16, TWH*17], facilitating the demand for the laborious object annotation.

In this study, we find the recently published arXiv paper [ZLVX16], close to our work. This study addresses the problem of MG classification from weakly labeled sets using MIL. Assuming that lesions occupy a small portion of the whole mammogram the authors add a sparsity constraint in the loss function enforcing the probability distribution of patches to be sparse (mostly negative, with zero probability). This study uses the CNN for representation of the whole mammogram and in order to use a pretrained network (on ImageNet) they downsize the large MG images by factor 7-14 on each side to reach 224$\times$224 size. The harsh downsizing action causes a significant loss of information in the mammogram. It is well known that malignant lesions often appear as masses or micro-calcifications and can be as small as 50$\times$50 pixels (cf. statistics on INbreast data set in [ZLVX16]). Yet, [ZLVX16] report a significantly high ROC-AUC of 0.859 on INbreast dataset [MAD*12]. However, this average AUC is based on 5-fold cross validation, without declaring the partition regime. Patient-wise partition of the folds is necessary to avoid images from the same patient to appear in the training and testing sets, otherwise contamination of the test set yields optimistic performance as a result of an overfit. In this study we use a *patch-based* approach with a max-pooling loss function, resulting localization in full resolution. Our post processing, employs top scored patches in full resolution and the test augmentations further contributes in improved results. Tang *et al.* [TWH*17] argue patch-based CNN advantages over plain CNN, reporting state-of-the-art results in weakly supervised classification and detection on natural images. We further conduct extensive experimental evaluations on a large data set ($\times$5 and $\times$8 of those used in [ZLVX16] and [HK16] respectively) and two different data splits, showing also localization results and accuracy.

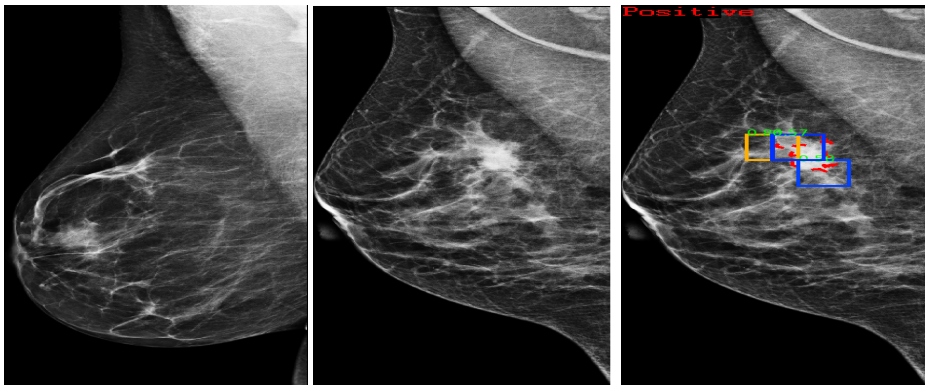
In this paper, we propose the following major contributions: (1) We present a novel patch-based deep multiple instance CNN for classification of a mammograms according to the severity of its internal findings, while trained on images with only global labels (without localization of the lesions). (2) Our approach further suggests detection of the lesions in full resolution. (3) We report results on a large multi-center mammogram data set as well as public data.

Our method further suggests the following advantages: (1) Decomposing the images into patches allows direct change of the local scale without the need to alter the CNN internal parameters or image resize. (2) Our model is insensitive to the image size and the number of the patches extracted from the image. So there is no need to warp the image to a fixed size, which often causes the distortion of the image and the lesion shape; (3) The patch based approach allows processing of non-rectangular regions in the image by masking of certain areas, with simply excluding patches from the collection. (4) A combined pre-trained CNN allows training on small data sets while shortening the training duration, since only the fully connected layers are then trained.

Weakly labeled approach using data sets tagged only at the global image level are particularly useful in medical image analysis, where the annotations often require expensive and time-consuming clinical expertise.

2. Methodology

Mammography frequently requires a binary classification task, separating normal cases from the rest or from those with suspiciously malignant findings. In traditional supervised learning, training data is given in pairs $\{(x_i, y_i)\}_{i=1}^{N}$, where x_i denotes the input image or features and $y_i \in \{y_+, y_-\}$ the label *e.g.* normal or benign versus malignant class label. Thus, the goal of supervised learning is to train a classifier $h : x \to y$ that will accurately predict the label y of a novel instance x. In the context of deep learning, the classifier becomes a multi-layer neural network, possibly convolutional, with the ability to create efficient feature representations of the global image data. However, in mammograms, discriminative information generally comes from relatively small local regions that eventually determine the global image label.



(a) Negative-Normal (b) Positive-Highly Suspicious as Malignant (c) Method Result

Figure 1: *Illustration of weakly labeled classification task. (a) Normal MG, (b) MG containing a malignant tumor, (c) the same image correctly classified as malignant overlaid by the radiologist annotation in dashed red contour (used only for validation). Our system alerts are shown by a heat colored bounding boxes. Lesions can be relatively small with respect to the whole image and occluded under the parenchymal tissues.*

Multiple instance learning is based on two basic ingredients, *bags* and *instances*. Bags can contain variable number of instances. In MIL, bag labels are given, and each bag consists of a collection of instances, with a-priori unknown labels (in training). MIL has two constraints: 1) if a bag is positive, at least one instance in the bag should be positive; 2) if a bag is negative, all instances in the bag should be negative. Images are typically represented by bags and patches as instances. Let the training set Λ consist of pairs of bags and their associated labels $\{(X_i,Y_i)\}_{i=1}^{N}$, where $X_i = \{x_{ij}\}_{j=1}^{m_i}$. Here, X_i denotes the bag, representing the i-th image with its label Y_i while x_{ij} presents the j-th patch in image i referred as instance. The goal of MIL is to classify unseen bags or instances based solely on the bag labels in the training stage. Thus, we assume that instance labels $y_{ij} \in \{y_+, y_-\}$ can exist for each instance, but are not known during training. The MIL assumption can then be satisfied by:

$$Y_i = \max_j (y_{ij}). \tag{1}$$

Considering the recent successes achieved by deep learning, it seems a natural choice to employ deep representations instead of shallow models. We use a deep CNN as our architecture for learning visual representation of patches, coupled with multiple instance learning to obtain a global representation. While the CNN architecture is similar to conventional supervised learning networks, the loss function must be adapted. In order to discriminate between classes, most classification frameworks involve cross entropy loss. However, in our binary scenario, negative patches are also present in positive images. Since they would obtain high probabilities even in positive scans, no separation between the two classes can be obtained. Thus, the cross entropy loss is modified to its MIL version, which defines a log likelihood loss by:

$$\mathcal{L}(\theta) = \sum_{\substack{X_i \in \Lambda \\ Y_i = y_+}} \log\left(\max_{x_{ij} \in X_i} \mathbb{P}(y_+|x_{ij},\theta)\right) + \sum_{\substack{X_i \in \Lambda \\ Y_i = y_-}} \log\left(1 - \max_{x_{ij} \in X_i} (\mathbb{P}(y_+|x_{ij},\theta)\right), \tag{2}$$

with $\mathbb{P}(y_+|x_{ij},\theta)$ as the probability that the local patch x_{ij} is classified as positive, based on the neural network coefficients θ. The two terms define the loss over the positive and negative sets respectively. Note that local patches (instances) has a-priori no label. The patch labels are assigned during the training process to optimize the bag classification.

There are areas in the MG image that carry no information, such as the breast exterior (see the black region in Fig. 2) or certain anatomical areas such as the skin or pectoral muscle, as shown in Fig. 2. In this approach we can use a modified probability $\omega(x_{ij})\mathbb{P}(y_+|x_{ij},\theta)$ where each patch probability is coupled with a probabilistic geometric prior $\omega(x_{ij})$ to alter the weight of patches in certain areas. One possible choice is to use some notion of distance such as:

$$\omega(x_{ij}) = 1 - \frac{\mathcal{A}(x_{ij} \cap \mathcal{B})}{\mathcal{A}(x_{ij})} \tag{3}$$

where $\mathcal{A}(x_{ij})$ denotes the area of patch x_{ij} and $\mathcal{B}$ the determined restricted area. We recognized two types of such areas. The first is the breast exterior and outline (skin). Thus patches intersecting breast outline were excluded from the "bag". The second excluded area was the pectoral muscle. Considering the pectoral muscle resulted inferior performance particularly due to existence of lymph nodes having a similar appearance as masses. We therefore disregard this area and consider it's further analysis in a future work. Note that findings near the restricted areas are still considered due to the overlapped patches. In the current experimental tests we used

a binary weighting:

$$\omega(x_{ij}) = \begin{cases} 1, & x_{ij} \cap \mathcal{B} = \emptyset \\ 0, & x_{ij} \cap \mathcal{B} \neq \emptyset, \end{cases} \quad (4)$$

This weighting and masking paradigm allows processing of non-rectangular regions in contrast to previous methods of [HK16, ZLVX16]. The breast contour was found using a global object pre-

Figure 2: *Excluding areas from the process, such as the breast exterior separated by the white curve and the pectoral muscle bounded by the blue curve and the top right corner.*

serving threshold, and the pectoral muscle by a dynamic programming that connected points of high gradient.

The new loss function satisfies a multi-instance learning criterion. Yet, this loss function is not differentiable. We therefore use a surrogate function by first sorting the patches according to their scores (*i.e.,* positive probability), then choosing the patch with the maximum score as input to a standard cross entropy function. This instance, presenting the most *discriminative* patch represents the whole image and is used for back propagation and update of the coefficients θ. Note that the suggested MIL paradigm is insensitive to the image size, since bags can contain a different number of instances.

3. Network Architecture

We start with breast segmentation, separating it from the background. The breast interior is then decomposed into fixed size patches with overlap. We use deep CNN as our framework for multiple instance learning. As mammograms are captured in high resolution to allow detection of small size findings, we extract our patches from the original image without downsampling. Due to the relatively small training data set, we employ a two-stage deep neural network architecture as depicted in Fig. 3. As the first stage we opted for a transfer learning approach by using the pretrained VGG-M network [CSVZ14a], trained on the ImageNet data set. In our model, we extract CNN codes from the last hidden layer as 4096D feature vector. This CNN network is followed by a refining fully connected neural net modified to comply with the MIL paradigm. The refining neural net includes three fully connected layers and is trained from scratch according to the suggested MIL loss function (2). Note that with a sufficiently large data set the network can be trained end-to-end for optimal image representation.

We use 224×224 local patches on the original scan (approx. 15 $\times$ 15 $[mm]$) as input to our pre-trained CNN, with 50% overlap. The VGG-M pretrained convolutional network is provided by the visual geometry group (VGG) of the University of Oxford, trained on the ImageNet data set [CSVZ14b]. This network consists of five convolution layers, three fully connected layers and a softmax layer with 1000 output nodes. In order to fit the image patches into the network, they were first transformed to 8 bit and, replicated to 3 RGB channels. In our model we use the output of the last fully connected layer (full7 layer-20) as a 4096D feature vector. Note that the VGG stage is fixed, namely the CNN weighs are frozen and not updated during the training. This strategy allows the patch feature vectors to be computed only once, prior to training, significantly reducing the computation cost as the refining network includes only three fully connected layers. We used Rectified Linear Units (ReLUs) as non-linear layers in the refined NN. The optimization step is obtained using momentum stochastic gradient descent.

To enlarge and **balance** the training set, we used augmentations by adding, rotations $7 \times 45^{\circ}$, flips and 6 shifts. The process is initialized with random labels over patches. The networks is then trained using stochastic gradient descent solver with a mini-batch size of 70-256 varied according to the data set (without batch normalization). We used a dynamic learning rate of $[0.5, 3.5] \times 10^{-3}$, momentum of 0.9 and weight decay of 10^{-3} and 10^{-4} for two different data sets. The stopping criterion was set to 20-30 epochs according to a validation set. We further post-process the results using 15 augmentations on the test patches as detailed above, then averaging over the top $K = 4$ instance scores to reach the image level probability. This post-processing improved outlier rejection and reduced the false positives.

The average training time was around 3 hours on NVIDIA-GTX Titan GPU (12GB), for 2,000 images of approximately $3K \times 1.5K$ size, in MatConvNet framework. The representation process (basically feed-forward through the VGG-CNN), took around 1.5 hours.

4. Experiments

We conducted our first experiment on a large multi-center hospital data set referred as **IMG**. The data set consists of 2,500 full-filed digital mammograms (FFDM) from a BI-RADS distribution of 1317, 662, 333 and 47, 141 corresponding to findings in the images associated with maximum BI-RADS 1,2,3 and 4,5 respectively. The mammograms contain various findings such as masses, macro and micro-calcifications. In our first test scenario we split the mammograms into the following two labels, BI-RADS 4,5,6 as positive (98 cases) and BI-RADS 1,2,3 as negative (780 cases). We included all types of suspiciously malignant abnormalities into the positive class to distinguish between any severe abnormality from BI-RADS 4,5 and normal images (BI-RADS:1) as well as the most likely) benign findings (BI-RADS:2 & 3). This data split raises a particular challenge as the model has to discriminate between images with a very similar types of lesions, such as malignant versus benign masses or different types of micro-calcifications, often ambiguous even for expert radiologists. Our second test bed includes the INbreast (INB) publicly available FFDM data set [MAD*12]. This relatively small data set includes 410 mammograms from 116 cases. We conducted the same split on the INbreast images, and obtained 100 positive and 310 negative mammograms. We refer to this test scenario as TS-1, and further test our model on a different

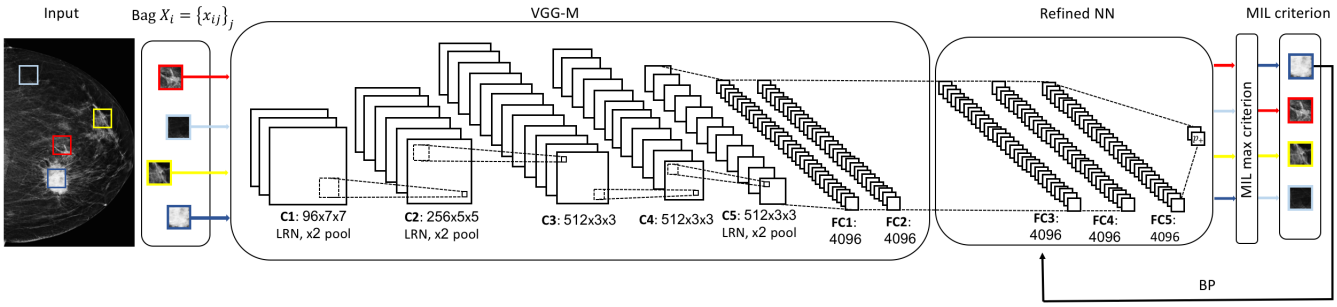


Figure 3: *The proposed deep multiple instance learning network architecture. After the preprocessing stage of breast segmentation, the image is decomposed into patches (instances). Each image patch is then run through a pretrained network to yield a 4096D representation vector. The refined NN then associates a class score to each feature vector (patch) according to its learned discrimination power. The probabilities are then processed to aggregate a final probability for the whole image (bag). Classification is determined according to the patch with the highest positive probability, whereas detection is obtained by selecting the patches with the highest positive predictive probabilities.*

scenario with a data split of BI-RADS 1 versus Rest illustrating a use case where the system alerts for any abnormality (even benign). We refer to this test scenario as TS-2.

We carried out our performance assessment with 5 fold **patient-wise** cross-validation. At each train and test iteration, all the images from the patient under testing were strictly excluded from the training set, to avoid data contamination and over-fitting.

4.1. Classification Results

Table 1 summarizes the experimental results for the classification task. We use the area under the ROC curve (AUC) measure for performance assessment due to the high imbalance distribution between classes in the data sets. Only 7.5% of mammograms in the IMG set are positive in TS-1. Our deep MIL model produces an average AUC of 0.831 ± 0.044 on the IMG data set in test scenario TS-1. Further analysis shows that on average, 48% of the false positives are from BI-RADS 2 & 3 categories. This shows that many network errors may be associated with wrong classification of masses and calcifications, which often pose a challenge even for expert radiologists. Testing on the small data set of INB resulted in a lower AUC of 0.722 ± 0.089. The lower AUC associated with high STD in INB reflects the influence of the small data size, on learning capability and validation. Note that for the commonly used 5-fold cross validation, there are approximately 16 positive images in each fold (only about 8 patients).

We further compare our method to two recent works and one baseline that reflects the impact of downsizing the images. The first reference [HK16] called Self-TL, uses a weakly labeled data set and a recently published method of ResNet MG [DCB17], employing a *fully annotated* image set (*i.e.,* including finding annotations). The latter ResNet MG presents results on INB in the same test scenario, TS-1. While Self-TL tests are similar to our TS-2, the data set is different, using digitally scanned MG. The third comparison from [SBAK17], conducted on the INB, comprises of a naïve transfer learning strategy (referred as Naïve-TL), in which the image (after cropping the area of the breast) is resized to 224×224 pixels. The breast image is then run through an off-the-shelf deep CNN model, trained on the Imagenet dataset (employing the same VGG-M model as used in this paper), obtaining from it a 4096 long representation vector for the entire image. These features are then used to train an SVM in order to classify them as positive or negatives. Note that this reference demonstrates the impact of image significant downsizing on the AUC.

The Self-TL method [HK16] yields an AUC measure of 0.675, a significantly lower performance on a similar size data set as INB, yet on a scanned MG set. With respect to the fully supervised learning method of ResNet MG tested on INB, our model achieves comparable result on the same data set when considering a single MG, but without requiring local annotations. The Naïve-TL demonstrates the significant impact of strongly downsizing the image resulting AUC of 0.602 on INB (see Table 1). Note that our model is further capable of distinguishing between different types of abnormalities such as micro-calcifications which can appear in both classes in TS-1.

In Fig. 4 we present the resulting ROC curves for our two test scenarios. For TS-1 with highly probable malignant MG classification, we obtain specificity of 60% @ 87% sensitivity or specificity of 40% @ 96% sensitivity. Similar work points for BI-RADS 1 vs. Rest shows specificity 60% @ 79% sensitivity or specificity 20% @ 96% sensitivity.

4.2. Localization Results

While the training starts with unlabeled instances, the final model classifies each instance/patch according to its discrimination power in separating positive and negative bags. We can use the instance score to visualize the discriminating regions that can be referred as abnormalities. Fig. 5 shows several true positive results from our deep MIL model on both IMG and INB data sets for TS-1. These results show the successful localization of the malignant findings in the image, whether it is a tumor or micro-calcification clusters.

Fig. 6 shows the visualization for TS-2. Results are shown with

Methodology	Labeling	TS	Data set	Type	# Images	Average AUC
Proposed Deep MIL	Weakly	(1)	IMG	FFDM	2,500	0.831 ± 0.044
Proposed Deep MIL	Weakly	(1)	INB	FFDM	410	0.722 ± 0.089
Proposed Deep MIL	Weakly	(2)	IMG	FFDM	2,500	0.817 ± 0.031
Proposed Deep MIL	Weakly	(2)	INB	FFDM	410	0.790 ± 0.093
Self-TL [HK16]	Weak	(2)	MIAS	Scanned	322	0.675
Naïve-TL from [SBAK17]	Weak	(1)	INB	FFDM	410	0.602
ResNet MG [DCB17]	Fully	(1)	INB	FFDM	410	0.740 ± 0.020 [1]

Table 1: Binary classification performance measured by average ROC-AUC for our approach in two different test scenarios (TS-1,2). Methods are differentiated by the type of labeling (Weakly vs. Fully), source and the size (# images) of the data set. The results shown depict two different test scenarios - TS-1: BIRADS 1,2,3 vs. 4,5,6 and TS-2: BIRADS 1 vs. Rest. For comparison, we show three relevant methods in the literature. Note that our weakly labeled model obtains comparable performance to the fully supervised method in [DCB17] on the same data set and same test scenario.

(1) Reaching AUC 0.8 when utilizing multiple views

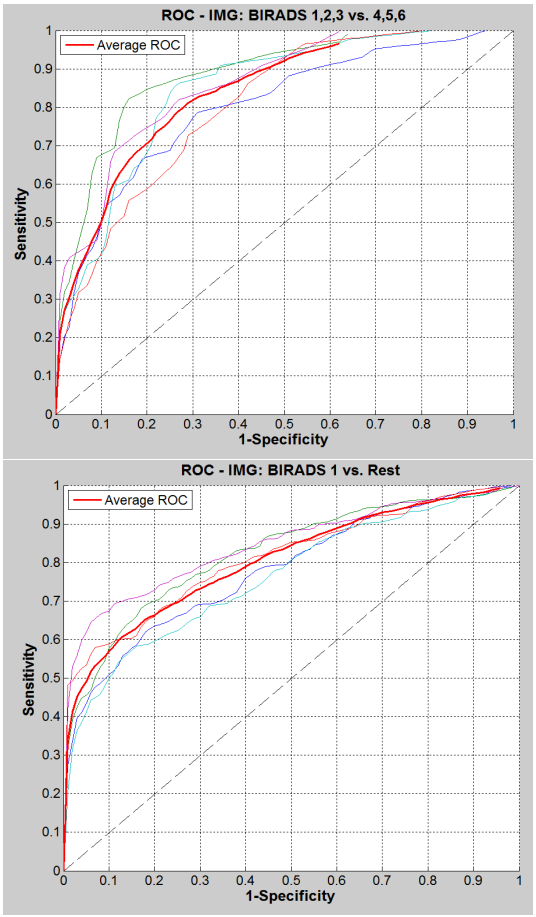


Figure 4: *ROC curves for our two test scenarios, over 5-fold cross validation. The thin plots describe the ROC for each fold while the wide red curve stands for the average ROC.*

the corresponding raw images to allow observation of the subtle findings captured by our method.

Lesions in our data set present a large scale variability of over 10 scale factor. Yet our patches are at fixed size and aim to alert for a suspicious finding, rather than exact segmentation of the lesion. Therefore, we use a less strict measure for localization than standard intersection over union. Considering a symmetric overlap ratio allows a small patch within a large mass be determined as true. Consequently, we would like to allow an extremely large finding to be covered by a single or several patches. Considering all top $K = 5$ patches, those having over 50% overlap with a true finding (or symmetrically if 50% of the lesion is covered by a patch) are considered as true positive patches, for localization. Accordingly, two false-positive measures are defined for localization derived from patches with an intersection ratio below 50%. The first measure F_D is the average false positive (detection) per-image (FPPI) in TP class and F_T, commonly used in the literature presenting the average false positive detection per image with respect to all the images in the cohort. At a work point of $F_D = 1$ FPPI our model for IMG data set in TS-1 yields an average recall rate of $R = 0.76$ @ $F_T = 0.48$. This means that on 76% of TP images we localize at least one lesion accurately, while keeping the total FPPI below 0.5.

The proposed framework allows scoring the patches according to the positive probability. The highly scored patches present the discriminative regions in the image and indicate the location of the abnormalities. The localization is an important feature allowing the analysis of the results or helping the user in understanding the system outcome. Note that we obtain this localization without having any local labels in the training set.

5. Summary

In this study, we propose a novel framework for classification of mammograms and detection of abnormalities with no local annotations available in the data set. Different use cases can be defined according to the data split in the training stage. One common classification scenario is according to the severity of the findings as

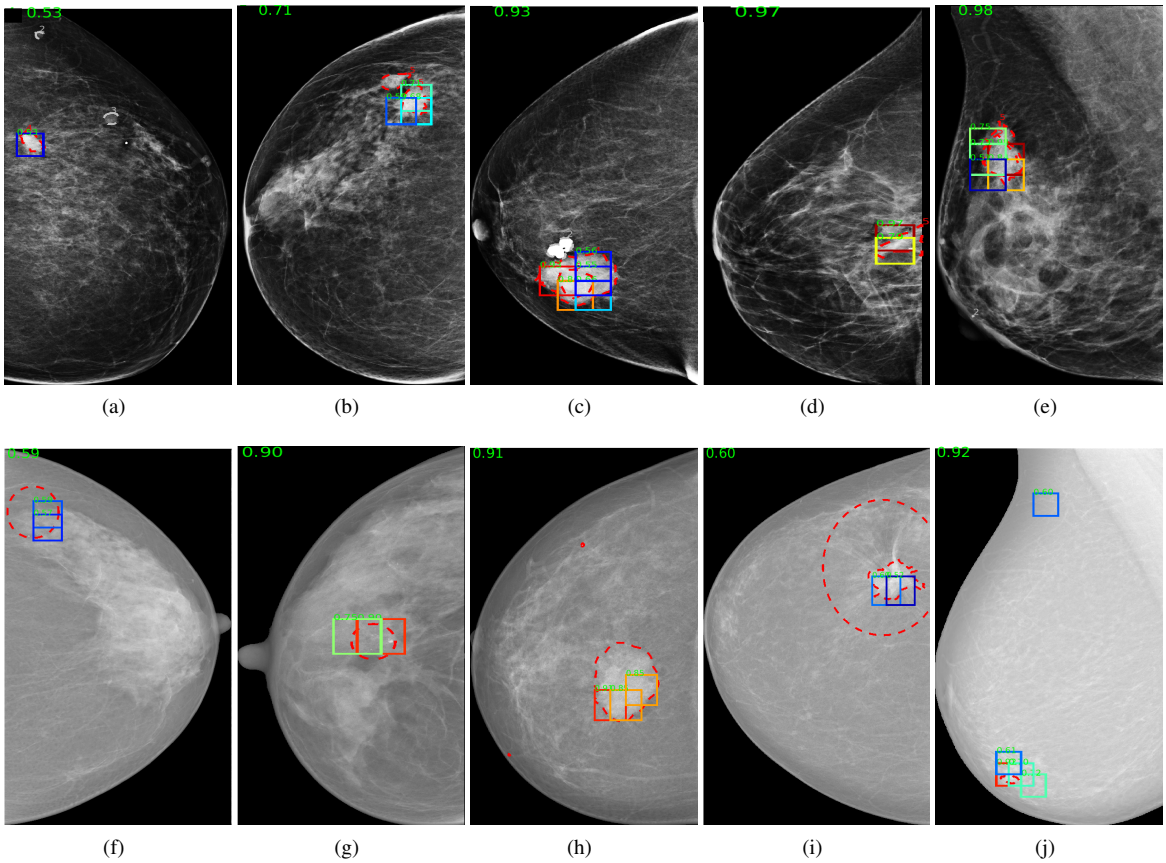


Figure 5: *Several true positive results from IMG (top-row) and INB (bottom-row) for test scenario-1, BI-RADS 1,2,3 vs. 4,5,6. Ground truth annotation is depicted in dashed red contour while our top instance predictions are overlaid. Jet color coding denotes positive probability from blue $p = 0.5$ to red $p = 1$. Note the overlap of our predicted patches with the ground truth. The overlaid numeric value at top left indicates the positive predictive value for the whole image. Note the ground-truth and instance predictions overlap as well as the benign macro-calcification in (c) correctly ignored. Image (j) contains a false-positive instance.*

performed by radiologists. Successful results in this challenging task may lead to strong implications in the field, since common supervised methods rely on finely annotated data that require the location and often delineation of findings in the image. Our proposed method classifies mammograms by detecting discriminative local information contained in patches, through a deep neural network with the multiple instance learning paradigm. Our framework jointly learns both the classifier and localization, which can be used as an analysis tool or for user explanation justifying the system classification decision. We tested our method on two FFDM data sets for mammogram classification and in two different test scenarios. The results are promising and come close to the existing fully labeled methods while requiring only a global tag over the image class. Consequently, it has a clear advantage for training on large data sets. The suggested CNN-based method eliminates the need for handcrafted features, and allows transferring the method to new modalities and organs with minimal overhead.

References

[BAABKH17] BEN-ARI R., AKSELROD-BALLIN A., KARLINSKY L., HASHOUL S.: Domain specific convolutional neural nets for detection of architectural distortion in mammograms. In *IEEE International Symposium on Biomedical Imaging* (2017). 1

[BV16] BILEN H., VEDALDI A.: Weakly supervised deep detection networks. In *CVPR* (2016). 2

[CC05] CADY B., CHUNG M.: Mammographic screening: no longer controversial. *American journal of clinical oncology 28*, 1 (2005), 1–4. 1

[CGGS13] CIREŞAN D. C., GIUSTI A., GAMBARDELLA L. M., SCHMIDHUBER J.: Mitosis detection in breast cancer histology images with deep neural networks. In *International Conference on Medical Image Computing and Computer-assisted Intervention* (2013), Springer, pp. 411–418. 1

[CSVZ14a] CHATFIELD K., SIMONYAN K., VEDALDI A., ZISSERMAN A.: Return of the devil in the details: Delving deep into convolutional nets. In *British Machine Vision Conference* (2014). 4

[CSVZ14b] CHATFIELD K., SIMONYAN K., VEDALDI A., ZISSERMAN A.: Return of the devil in the details: Delving deep into convolutional nets. In *British Machine Vision Conference* (2014). 4

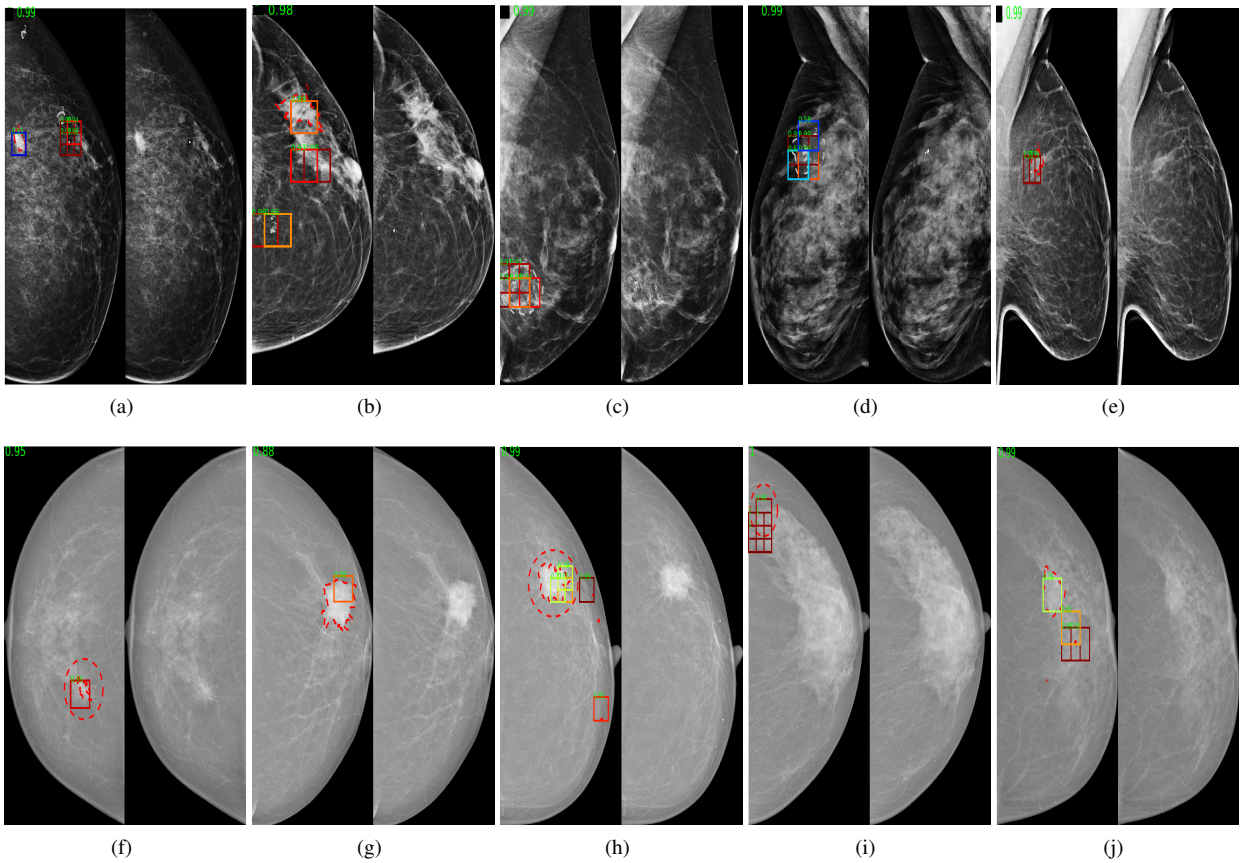


Figure 6: *Several true positive results from IMG (top-row) and INB (bottom-row) for test scenario-2, BI-RADS 1 vs. Rest. Ground truth annotation is depicted in dashed red contour while our top instance predictions are overlaid. Jet color coding denotes positive probability from blue to red $p \in [0.5, 1]$. Overlaid images are shown with raw counterpart to allow better observation of the subtle findings. Note how the capture of micro-calcifications in (a) and (b) as well as the malignant tumor in (b). The alerted findings correspond to the data split and represent discrminative regions in the positive set.*

[DCB16] DHUNGEL N., CARNEIRO G., BRADLEY A. P.: The automated learning of deep features for breast mass classification from mammograms. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2016), Springer, pp. 106–114. 1

[DCB17] DHUNGLE N., CARNERIO G., BRADLEY A. P.: Fully automated classification of mammograms using deep residual neural networks. In *IEEE International Symposium on Biomedical Imaging* (2017). 1, 5, 6

[DLLP97] DIETTERICH T. G., LATHROP R. H., LOZANO-PÉREZ T.: Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence 89*, 1 (1997), 31–71. 2

[dOdCFS*15] DE OLIVEIRA F. S. S., DE CARVALHO FILHO A. O., SILVA A. C., DE PAIVA A. C., GATTASS M.: Classification of breast regions as mass and non-mass based on digital mammograms using taxonomic indexes and svm. *Computers in biology and medicine 57* (2015), 42–53. 1

[HK16] HWANG S., KIM H.-E.: Self-transfer learning for weakly supervised lesion localization. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)* (2016), Springer, pp. 239–246. 2, 4, 5, 6

[JBC*11] JEMAL A., BRAY F., CENTER M. M., FERLAY J., WARD E., FORMAN D.: Global cancer statistics. *CA: a cancer journal for clinicians 61*, 2 (2011), 69–90. 1

[JY15] JEN C.-C., YU S.-S.: Automatic detection of abnormal mammograms in mammographic images. *Expert Systems with Applications 42*, 6 (2015), 3048–3055. 1

[KPN*16] KALLENBERG M., PETERSEN K., NIELSEN M., NG A. Y., DIAO P., IGEL C., VACHON C. M., HOLLAND K., WINKEL R. R., KARSSEMEIJER N., ET AL.: Unsupervised deep learning applied to breast density segmentation and mammographic risk scoring. *IEEE transactions on medical imaging 35*, 5 (2016), 1322–1331. 1

[MAD*12] MOREIRA I. C., AMARAL I., DOMINGUES I., CARDOSO A., CARDOSO M. J., CARDOSO J. S.: Inbreast: toward a full-field digital mammographic database. *Academic radiology 19*, 2 (2012), 236–248. 2, 4

[MLP98] MARON O., LOZANO-PÉREZ T.: A framework for multiple-instance learning. *Advances in neural information processing systems* (1998), 570–576. 2

[OGA*08] OLIVEIRA J. E. D., GULD M., ARAUJO A., OTT B., DESERNO T. M.: Towards a standard reference database for computer-aided mammography. In *Proceedings of SPIE Medical Imaging* (2008). 2

[PND*14] PETERSEN K., NIELSEN M., DIAO P., KARSSEMEIJER N., LILLHOLM M.: Breast tissue segmentation and mammographic risk

scoring using deep learning. In *International Workshop on Digital Mammography* (2014), Springer, pp. 88–94. 1

[QLC*16] QUELLEC G., LAMARD M., COZIC M., COATRIEUX G., CAZUGUEL G.: Multiple-instance learning for anomaly detection in digital mammography. *IEEE transactions on medical imaging 35*, 7 (2016), 1604–1614. 2

[SBAK17] SULAM J., BEN-ARI R., KISILEV P.: Maximizing auc with deep learning for classification of imbalanced mammogram datasets. In *Eurographics Workshop on Visual Computing for Biology and Medicine* (2017). 5, 6

[SHLKR16] SUN M., HAN T. X., LIU M.-C., KHODAYARI-ROSTAMABAD A.: Multiple instance learning convolutional neural networks for object recognition. In *Proceedings of the International Conference on Pattern Recognition (ICPR)* (2016). 2

[SZ14] SIMONYAN K., ZISSERMAN A.: Very deep convolutional networks for large-scale image recognition. *CoRR abs/1409.1556* (2014). 1

[TWH*17] TANG P., WANG X., HUANG Z., BAI X., LIU W.: Deep patch learning for weakly supervised object classification and discovery. *Pattern Recognition* (2017). 2

[WYHY15] WU J., YU Y., HUANG C., YU K.: Deep multiple instance learning for image classification and auto-annotation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015), pp. 3460–3469. 2

[YCBL14] YOSINSKI J., CLUNE J., BENGIO Y., LIPSON H.: How transferable are features in deep neural networks? In *Advances in neural information processing systems* (2014), pp. 3320–3328. 1

[YZP*16] YAN Z., ZHAN Y., PENG Z., LIAO S., SHINAGAWA Y., ZHANG S., METAXAS D. N., ZHOU X. S.: Multi-instance deep learning: Discover discriminative local anatomies for bodypart recognition. *IEEE transactions on medical imaging 35*, 5 (2016), 1332–1343. 2

[ZLVX16] ZHU W., LOU Q., VANG Y. S., XIE X.: Deep multi-instance networks with sparse label assignment for whole mammogram classification. In *arXiv* (2016). 2, 4