

CLASSIFICATION AND DETECTION IN MAMMOGRAMS WITH WEAK SUPERVISION VIA DUAL BRANCH DEEP NEURAL NET

Ran Bakalo^{1,2} Rami Ben-Ari² Jacob Goldberger³

¹ Department of Computer Science, University of Haifa, Israel, ² IBM Research, Haifa, Israel

³ Faculty of Engineering, Bar Ilan University, Israel

ABSTRACT

The high cost of generating expert annotations, poses a strong limitation for supervised machine learning methods in medical imaging. Weakly supervised methods may provide a solution to this tangle. In this study, we propose a novel deep learning architecture for multi-class classification of mammograms according to the severity of their containing anomalies, having only a global tag over the image. The suggested scheme further allows localization of the different types of findings in full resolution. The new scheme contains a dual branch network that combines region-level classification with region ranking. We evaluate our method on a large multi-center mammography dataset including $\sim 3,000$ mammograms with various anomalies and demonstrate the advantages of the proposed method over a previous weakly-supervised strategy.

Index Terms— Mammography, deep learning, weakly supervision

1. INTRODUCTION

Screening mammograms (MG) are the first line of imaging for the early detection of breast cancer, that raise the survival rate, but place a massive workload on radiologists. Although MG provides a high resolution image, its analysis remains challenging, due to tissue overlaps, high variability between individual breast patterns, and subtle malignant findings.

A naive approach for classifying mammograms suggests training a detector from local annotations [1, 2], and then classifying the images according to the most severe detection. However, this requires instance level annotations for every finding in the image which is tedious, costly and impractical for massive datasets. A common technique to combat this problem is using weakly supervised paradigms, where labels are given only at the image-level (*e.g.* knowing a certain image contains a malignant finding but without localization). Some weakly supervised methods can further provide localization, as discriminative regions. This can be used as explanation behind the model decision particularly where the source of discrimination between the classes is a-priori unknown.

Previous weakly supervised approaches for mammogram classification [3, 4] used the popular multi-instance learning (MIL) concept. Hwang *et al.* [5] used a CNN with classification and localization branch with a linear combination loss. The shortcomings of these methods include: restriction to binary classification, processing of harshly downsized images [5, 4], implicit detection paradigm via max-pooling operation [3, 4] and localization at low resolution [5, 4]. Recent studies suggest that applying explicit *data-driven* detection in parallel to classification yields improved performance [6].

In this study we explore a new data-driven strategy for weakly supervised learning with a novel dual-branch deep CNN model. Our architecture is composed of two streams, acting on image regions, one for classification and the other for detection. The classification compares association of regions with certain class *e.g.* normal, benign or malignant. In the detection branch however, the scores of regions are compared to others, at each class. Hence, the classification branch predicts which class to associate to a region, whereas the detection branch selects which regions are more likely to contain a finding.

Our approach is inspired by [6] (applied on natural images) but differs and generalizes the existing methods by two main factors: 1) We don't use any unsupervised region proposal in our scheme as it is commonly unavailable in MG. 2) We change the architecture in [6] and extend the classification stream with additional class but without any detection counterpart, in order to handle images lacking any object (finding). The extension for handling images without objects is equivalent to normal radiology images without any pathology. The additional normal class changes the score distribution for regions, impacting the global score used for image classification. Similar to [7], we further connect between the branches by adding the information from the classification branch to guide the detection to the most relevant regions. Our model is capable of multi-class classification and detection that provides anomaly localization in full resolution. As such we compare our method with previous approaches in both classification and detection, showing improved performance.

2. MAMMOGRAM CLASSIFICATION ALGORITHM

Generally, radiologists divide regions in MG into three classes by severity order: normal tissues that can be viewed as background, lesions appearing to be benign and finally malignant findings. Our goal is to automatically classify the image into one of these classes. We suggest a novel weakly supervised and deep learning model composed of two branches as described in Fig. 1. The branches represent: (1) a *classification* that computes probabilities for classes *e.g.* malignant, benign with addition of a normal class for each image region, (2) a *detection* branch that compares all regions and ranks them according to the belief of containing a pathology. It is therefore a proxy to detection. The branches are then combined downstream to obtain an image-level decision for the presence of malignant and/or benign findings (see Fig. 1). In the following sections we describe in detail the building blocks of our scheme.

Region extraction. An image x , is first decomposed into m regions denoted by x_1, \dots, x_m . To this end, we used a sliding window of 224×224 overlapping regions (with 112×112 stride) in the breast excluding the axilla (which should be handled separately due to the different nature of the findings in this region). Through several layers of CNN each region is encoded into a feature representation. Due to the relatively small training dataset, we employ a two-stage method. In the first stage, we apply a transfer-learning approach by using a pretrained VGG128 network [8], trained on the ImageNet dataset [9]. In our model, we extract CNN codes from the last hidden layer such that $x_i \in \mathbb{R}^{128}$ is the feature vector representation of the i -th region (other VGG configurations and layers yielded inferior results). Then, we process each region separately by a fully connected (FC) layer (FC3 in Fig. 1).

Classification branch. Each region is classified (soft decision) as N , B or M , associated with healthy (normal) tissue, benign or malignant abnormality. To this end we use the following softmax layer (corresponding to lower branch in Fig. 1):

$$p_{\text{cls}}(c|x_i) = \frac{\exp(w_c^\top x_i)}{\sum_{d \in \{N, B, M\}} \exp(w_d^\top x_i)} \quad (1)$$

$$c \in \{N, B, M\}, \quad i = 1, \dots, m$$

such that w_N, w_B and w_M are the parameters of the respective classifier. Note that the same classification parameters are used across all the regions in the image.

Region selection. It makes sense to connect between the decision processes of classification and detection. For example, if a region is strongly classified as M we would like that the malignancy detector will favor this region. We therefore impose a soft alignment between the branches by making the classification branch selecting regions for detection. Note that the classification branch includes a separate normal class, thus allowing to filter out normal regions from the detection branch

(by assigning them low probabilities). We formalize this intuition by a region selection scheme. For each abnormality class in the detection branch we consider the corresponding k top scoring regions, removing the regions that obtained low classification scores. As the number of selected regions we chose $k = 10$. This implicit region selection scheme guides the detector to focus on the most likely abnormal regions in one hand and forces the network to improve classification in other hand (due to loss amplification on those errors). For normal breast images, where all regions are of class N , erroneously classified regions act as *hard negative mining*.

Detection branch. In parallel, we compute the relevance of each region as an abnormality by ranking (upper branch in Fig. 1). There is a separate ranking for each class; in this study one for malignant and the other for benign. Since the prevalent normal regions are not considered as abnormality, these regions are treated as “background” and therefore not associated with a separate detection class. The detection result is therefore a distribution over all the regions in the image:

$$p_{\text{det}}^c(i|x) = \frac{h_c(i) \exp(u_c^\top x_i)}{\sum_{j=1}^m h_c(j) \exp(u_c^\top x_j)} \quad (2)$$

$$c \in \{B, M\}, \quad i = 1, \dots, m$$

such that u_B and u_M are the parameter-set of the abnormality detectors and $h_c(i)$ is a k -hot vector indicating the k -top scoring regions in the associated classification branch. Note that for certain class c , Eq. (2) is equivalent to ranking of i -th region in image x with respect to other regions.

Image level decision. Given the region-level classification results and the region detection distribution, we can now evaluate the image-level class. Let (y_M, y_B) be a binary tuple indicator whether an image contains a malignant and/or benign finding, respectively. Note that this type of tuple labeling allows tagging images of class N by $(0, 0)$ and those with both M and B findings by $(1, 1)$. The posterior distributions of y_M and y_B are obtained as a weighted average of the local decisions:

$$p(y_c = 1|x) = \sum_{i=1}^m p_{\text{det}}^c(i|x) p_{\text{cls}}(c|x_i), \quad c \in \{B, M\}. \quad (3)$$

Since in many medical applications such as mammography the most prevalent cases are normal without any finding, we extended the method in [6] by adding a normal (N) classifier in the classification branch. This is a novel generalization to previous modeling in [6] where each object class is associated with a corresponding detector. Normal images in our model are then discriminated by having low probability for all types of abnormalities. The addition of N classifier further reduces false positives in detection (localization) as normal regions gain low scores (instead of uniform scores over all classes of abnormalities). The probability for an image to be normal can be obtained via the joint probability $p(y_M = 0, y_B = 0|x)$.

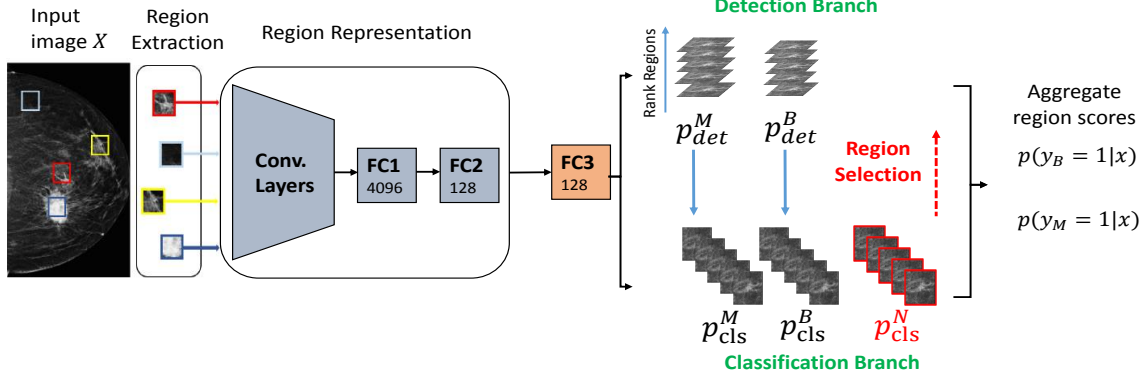


Fig. 1: Architecture overview. The novel elements are in red. Our new scheme has an additional class in the classification branch (p_{cls}^N) with no associated detection. This allows filtering of normal regions to obtain a low global score for a normal breast as desired. The region selection further boosts performance.

Training. The network output are soft decisions for each train image $x(t)$ as containing any type of abnormality *e.g.* M or B , represented by the tuple label $(y_M(t), y_B(t))$. We set the following cross-entropy loss function for the network training:

$$L(\theta) = - \sum_{c \in \{M, B\}} \sum_{t=1}^n \log p(y_c(t) | x(t); \theta) \quad (4)$$

such that θ is the parameter-set of the model and the probability defined in Eq. (3). Our infrastructure and parameter setting include TensorFlow framework using Adam optimizer with learning rate 10^{-4} , dropout, l_2 -regularization and a batch-size of 256 images (with 200 regions in average per image). To enlarge and balance the training set, we used augmentations by rotations, flips and image shifts.

3. EXPERIMENTAL RESULTS

Dataset. For evaluation of our method we used a large multi-center full field digital mammography (FFDM) screening dataset acquired from various devices (with approximately $3K \times 1.5K$ image size). From this proprietary dataset we excluded images containing artifacts such as metal clips, skin markers, etc., as well as large foreign bodies (pacemakers, implants, etc.). Otherwise, the images contain a wide variation in terms of anatomical differences, pathology (including benign and malignant cases) and breast densities that corresponds to what is typically found in screening clinics. In terms of global image BI-RADS (Breast Image Reporting and Diagnostics System), we had 350, 2364, 146 and 107 corresponding to BI-RADS 1,2,4 and 5 respectively. Global BI-RADS were set by the most severe finding in the image, according to clinical guidelines. Mammograms with global BI-RADS 3 were excluded since these intermediate BI-RADS are commonly assigned based on other modalities (*e.g.* US) and comparison to prior mammograms [10],

which are not available in our test case. However, our data set includes BI-RADS 3 findings that were not the most severe ones in the image. We split the mammograms into the following three global labels: BI-RADS 4 & 5 defined as malignant (M), BI-RAD 2 as benign (B) and BI-RAD 1 as normal (N).

Our evaluation is based on 5 fold **patient-wise** cross-validation, where folds were spitted by patient IDs. We also used the INbreast (INB) public FFDM dataset [11] for our evaluation. We split it into 100 positive (global BI-RADS 4,5,6) and 310 negative (global BI-RADS 1,2,3) mammograms. As this is a small data set, we conducted a random patient split on INB with 50% for train and 50% for test.

Evaluation. We compare our proposed model (Cls-Det-RS) to the previously published method of [3] (Max-region) and the DB-Baseline presenting an equivalent dual-branch approach of [6]. For evaluation, we present our results on two binary classification tasks, frequently used in MG, by grouping two “nearby” classes. To this end, we used $p(y_M = 1|x)$ scoring for M vs. $B \cup N$ (M vs. BN) and $\max\{p(y_M = 1|x), p(y_B = 1|x)\}$ scoring for $M \cup B$ vs. N (MB vs. N).

In addition to AUROC we also report two more practical measures as used in [12]. The partial-AUC ratio (pAUCR), associated with the ratio of the area under the ROC curve in a high sensitivity range ($[0.8, 1]$), representing the AUROC at a more relevant domain for clinicians. Also the specificity extracted from the ROC curve at sensitivity 0.85 representing an average operation point (OP) for expert radiologists, as reported in [13]. Table 1 presents the binary classification performance in two classification scenarios of M vs. BN and MB vs. N. Our method outperformed both the DB-Baseline and Max-region [3] in all measures and both classification scenarios. There was an improvement of 10-17% on pAUCR and 8-11% on specificity at sensitivity 0.85 ($p < 0.001$). In the case of MB vs. N we obtain 6-13% higher pAUCR and an 8-14% higher specificity at 0.85 sensitivity ($p < 0.001$). In terms of AUROC, the results showed an improvement of 2-4% with respect to the previous Max-region and the DB-

Baseline. We obtained a similar trend at breast level analysis (using maximum score between views). Train and test on the small public data set of INB yielded AUROC of 0.73. This is comparable to reported results in the literature, and close to the fully-supervised results on single view in [14] with AUROC of 0.74. Our region based error analysis indicates that the majority of our errors are between B and M class as it is also the case for breast radiologists.

Method	AUROC	pAUCR	OP
M vs. BN weak supervision			
DB-Baseline [6]	.709±.020	.251±.05	.37
Max-region [3]	.699±.047	.235±.10	.36
Cls-Det-RS	.728±.036	.275±.10	.40
MB vs. N weak supervision			
DB-Baseline [6]	.826±.01	.347±.03	.51
Max-region [3]	.817±.02	.323±.07	.48
Cls-Det-RS	.841±.02	.367±.05	.55

Table 1: Binary classification performance. OP: Specificity @ 0.85 sensitivity.

Our model further allows localization of abnormalities by highlighting the regions with top scores. We define this score as joint probability for detection (with $h_c \equiv 1$) and classification for each class:

$$d^c(x_i) = p_{cls}(c|x_i)p_{det}^c(i|x) \quad i \in \{1, \dots, m\} \quad (5)$$

Figure 2 shows several examples with localization in the test set, overlaid with radiologist annotations (used only for validation). As observed, the method is capable of separately highlighting benign and malignant lesions without having an instance level annotation. We further compare our method

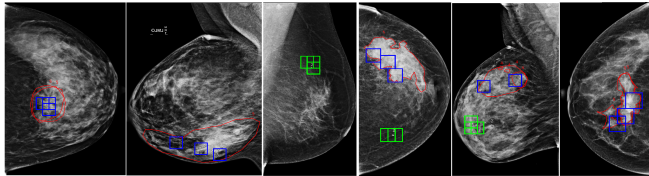


Fig. 2: Localization in classified images. Malignant and Benign lesions are annotated in red and gray respectively. Top 3 M (blue) and B (green) region scores are shown. Note the correlation between the radiologist annotation and the model prediction for each class.

performance in localization to Max-region and DB-Baseline. Due to large scale factor in lesion size we use the accuracy measure Intersection over Minimum (IoM) in [15]:

$$IoM(i, c) = \frac{|r_i \cap c|}{\min\{|r_i|, |c|\}}, \quad (6)$$

with c as the delineated domain and r_i as the extracted region. We define correct localization as $IoM \geq 0.5$. This measure allows explanation of an outcome when a specified region contains a true type of lesion or vice-versa. Note that the region size is relatively small and fixed (see Fig. 2). We present the FROC localization accuracy for class $c \in \{M, B\}$ using $d^c(x_i) \geq Threshold$. The recall rate R in the FROC is the fraction of images in the True-Positive set with at least one correct localization. The FROC curves in Fig. 3 show the superior performance of the proposed method in localization. Note that for M vs. BN improvement is more emphasized. This scenario sets a higher challenge since confusion between detection of M & B are penalized in FROC score.

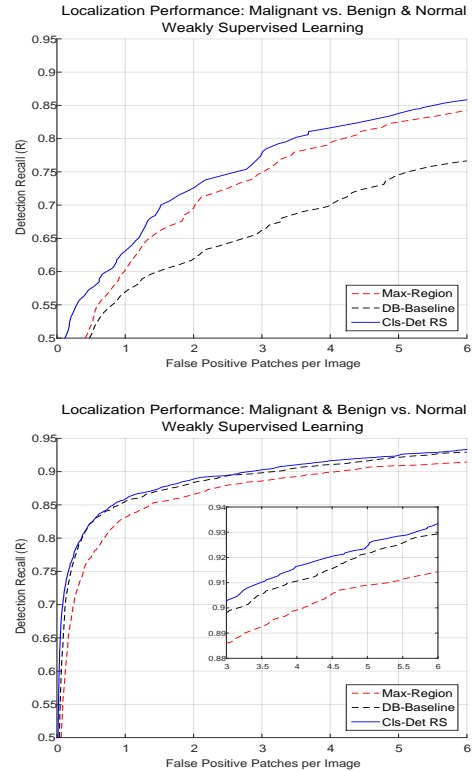


Fig. 3: Localization FROC at 0.85 sensitivity in classification. Note the superior performance (higher curve) of the proposed Cls-Det-RS method, particularly in M vs. BN.

To conclude, in this work, we propose a weakly supervised method for classification and detection of abnormalities in mammograms. We consider the classification of prevalent normal cases in MG as a separate class, lacking any abnormality. Our future work includes the aggregation of extra views with bilateral and prior images, into the weakly supervised setting. These additional features together with our inexpensively labeled data and explanatory model, will close the gap towards a practical tool for radiologists.

4. REFERENCES

- [1] Dezso Ribli, Anna Horváth, Zsuzsa Unger, Péter Pollner, and István Csabai, “Detecting and classifying lesions in mammograms with deep learning,” *CoRR*, vol. abs/1707.08401, 2017.
- [2] William Lotter, Greg Sorensen, and David Cox, “A multi-scale CNN and curriculum learning strategy for mammogram classification,” *CoRR*, vol. abs/1707.06978, 2017.
- [3] Yoni Choukroun, Ran Bakalo, Rami Ben-Ari, Ayelet Akselrod-Ballin, Ella Barkan, and Pavel Kisilev, “Mammogram Classification and Abnormality Detection from Nonlocal Labels using Deep Multiple Instance Neural Network,” in *Eurographics Workshop on Visual Computing for Biology and Medicine*. 2017, The Eurographics Association.
- [4] Wentao Zhu, Qi Lou, Yeeleng Scott Vang, and Xiaohui Xie, “Deep multi-instance networks with sparse label assignment for whole mammogram classification,” in *MICCAI*, 2017, pp. 603–611.
- [5] Sangheum Hwang and Hyo-Eun Kim, “Self-transfer learning for weakly supervised lesion localization,” in *Medical Image Computing and Computer-Assisted Intervention - MICCAI*, 2016, pp. 239–246.
- [6] Hakan Bilen and Andrea Vedaldi, “Weakly supervised deep detection networks,” in *CVPR*, 2016, pp. 2846–2854.
- [7] Wenhui Jiang, Thuyen Ngo, B. S. Manjunath, Zhicheng Zhao, and Fei Su, “Optimizing region selection for weakly supervised object detection,” *CoRR*, vol. abs/1708.01723, 2017.
- [8] Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman, “Return of the devil in the details: Delving deep into convolutional nets,” in *BMVC*, 2014.
- [9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li, “Imagenet: A large-scale hierarchical image database,” in *CVPR*, 2009, pp. 248–255.
- [10] Aya Y. Michaels, Robyn L. Birdwell, Chris SungWon Chung, Elisabeth P. Frost, and Catherine S. Giess, “Assessment and management of challenging bi-rads category 3 mammographic lesions,” *RadioGraphics*, vol. 36, no. 5, pp. 1261–1272, 2016.
- [11] Inês C Moreira, Igor Amaral, Inês Domingues, António Cardoso, Maria João Cardoso, and Jaime S Cardoso, “Inbreast: toward a full-field digital mammographic database,” *Academic radiology*, vol. 19, no. 2, pp. 236–248, 2012.
- [12] “The Digital Mammography DREAM Challenge,” <https://www.synapse.org/#!/Synapse:syn4224222>.
- [13] Constance D. Lehman, Robert F. Arao, Brian L. Sprague, Janie M. Lee, Diana S. Buist, Karla Kerlikowske, Loise M. Henderson, Tracy Onega, Anna N. A. Tosteson, Garth H. Rauscher, and Diana L. Miglioretti, “National performance benchmarks for modern screening digital mammography: Update from Breast Cancer Surveillance Consortium,” *Radiology*, vol. 283, no. 1, pp. 59–69, 2017.
- [14] Neeraj Dhungle, Gustavo Carnerio, and Andrew P. Bradley, “Fully automated classification of mammograms using deep residual neural networks,” in *ISBI*, 2017.
- [15] Rami Ben-Ari, Ayelet Akselrod-Ballin, Leonid Karlinsky, and Sharbel Hashoul, “Domain specific convolutional neural nets for detection of architectural distortion in mammograms,” in *ISBI*, 2017.