Neurocomputing 421 (2021) 15–25

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Weakly and semi supervised detection in medical imaging via deep dual branch net

Ran Bakalo [a,b], Jacob Goldberger [c,*], Rami Ben-Ari [b]

[a] Department of Computer Science, University of Haifa, Israel
[b] IBM Research, Haifa, Israel
[c] Faculty of Engineering, Bar-Ilan University, Ramat-Gan 52900, Israel

ARTICLE INFO

Article history:
Received 6 June 2019
Revised 9 April 2020
Accepted 19 September 2020
Available online 30 September 2020
Communicated by Zidong Wang

Keywords:
Weakly supervised detection
Semi-supervised detection
Deep learning
Abnormality detection
Mammography
Breast radiology

ABSTRACT

This study presents a novel deep learning architecture for multi-class classification and localization of abnormalities in medical imaging illustrated through experiments on mammograms. The proposed network combines two learning branches. One branch is for region classification with a newly added normal-region class. Second branch is region detection branch for ranking regions relative to one another. Our method enables detection of abnormalities at full mammogram resolution for both weakly and semi-supervised settings. A novel objective function allows for the incorporation of local annotations into the model. We present the impact of our schemes on several performance measures for classification and localization, to evaluate the cost effectiveness of the lesion annotation effort. Our evaluation was primarily conducted over a large multi-center mammography dataset of ∼3,000 mammograms with various findings. The results for weakly supervised learning showed significant improvement compared to previous approaches. We show that the time consuming local annotations involved in supervised learning can be addressed by a weakly supervised method that can leverage a subset of locally annotated data. Weakly and semi-supervised methods coupled with detection can produce a cost effective and explainable model to be adopted by radiologists in the field.

1. Introduction

The most common type of cancer and the second leading cause of death in women is breast cancer [1]. Nearly 40 million mammography exams are performed on a yearly basis in the US alone. Screening mammograms (MG) are the first line of imaging for the early detection of breast cancer. These raise the survival rate, but place a massive workload on radiologists. Although mammography provides a high resolution image, its analysis remains challenging because of tissue overlaps, the high variability between individual breast patterns, subtle malignant findings (often less than 0.1% of the image area) and the high similarity between benign and malignant lesions. Suspicious lesions are often difficult to detect and classify, even by expert radiologists. Lesions can be relatively small with respect to the whole image and occluded in the parenchymal tissues.

A broad range of traditional machine learning classifiers have been developed for automatic diagnosis of specific findings such as masses and calcifications, and ultimately breast cancer [2,3]. Ultimately, diagnosis in mammograms is often dictated by the type of lesion found.

Our goal is building an automatic system that can jointly detect the lesion location (if it exists) and analyze the findings. This goal can be achieved by training a detector from local (often referred as instance) annotations [4,5], and then classifying the image according to the most severe finding in the image. However, in this type of supervised setting, training requires bounding-box annotations for every single abnormality. This setting is tedious, costly and impractical for large data sets. This problem is exacerbated in mammograms that can contain tens or hundreds of micro-calcifications spread throughout the breast. Having manual annotations further increases the likelihood of inconsistency in labeling due to a lack of consensus between radiologists [6] caused by ambiguous lesion boundaries. This problem is often resolved by having multiple annotators [7] that further escalates the workload.

In the weakly supervised paradigm, only global image-level tags are provided to train a classifier. Global image labels are easily available from retrospective clinical records often without the need for further clinician intervention. Weak supervision, however, provides no local information on the lesion location. In an era of growing demand for XAI (explainable AI), localization can shed light on

* Corresponding author.
E-mail address: jacob.goldberger@biu.ac.il (J. Goldberger).

https://doi.org/10.1016/j.neucom.2020.09.037
0925-2312/

the model reasoning for the image classification, and help foster trust among practitioners in the field. Hence, weakly supervised methods which also localize abnormalities provide high value especially in scenarios where the source of discrimination between the classes is a priori unknown.

In this study we address the acute problem of annotation and suggest a new network that can be trained on weakly labeled data and is capable of localizing the lesions at test time (perform detection), in full resolution. Our network architecture is composed of two branches (streams), one for classification and the other for detection. In the classification branch regions are classified to abnormality classes (*e.g.* benign or malignant) and a newly added *normal-region* class representing healthy tissues. In the detection branch the scores of all regions are ranked relative to one another, for each abnormality class (resulting in a distribution over regions per abnormality class). The classification branch classifies each region, whereas the detection branch selects which regions are more likely to contain a finding. The image class probability is then obtained by aggregation of the detection and classification probabilities for all regions in the image. The final abnormality probability is then increased when a suspicious finding is contained in one of the regions, similar to a radiologist's inspection work flow.

The main contributions of this work are as follows: 1. A dual branch deep learning architecture for joint image classification and region detection via region classification branch with a newly added normal-region class and, in parallel, region ranking branch. 2. A weakly-supervised learning method to train the proposed network. 3. A semi-supervised learning method to train the proposed network. Our method enables joint learning using weakly supervised data and additional fully supervised data with a novel region-level objective function on the branches' region-level probabilities.

Semi-supervised datasets combine globally labeled data with a small amount of data with explicit local annotations in addition to the global labels. In this work, we explore the problem of training the described network in weakly supervised and semi-supervised setups. The results of the proposed system are illustrated in Fig. 1.

We validate our method on a large FFDM dataset of nearly 3,000 mammograms as well as the public INBreast dataset [8]. Direct comparison of our method to previous works [9,10] and an ablation study shows that our model outperforms others in classification and, in particular, in detection.

A preliminary version of this work (with only weakly supervised setting) has been reported [11]. Our study include additional results analysis, ablation study and addition of a semi-supervised learning method.

2. Related work

Deep learning methods promise a breakthrough on assisting breast radiologists for early cancer detection in mammograms. However, the bottleneck for supervised methods in Big Data is the annotation workload which often requires expert clinicians/radiologists to delineate numerous benign as well as malignant findings in mammograms. Weakly supervised and semi-supervised methods are considered an affordable compromise to this tangle.

2.1. Weakly supervised detection

Weakly supervised detection methods in deep learning have attracted growing interest with the publish of the paper "Is localization for free?" [12] that addressed the tedious task of local annotations in images [10,13]. Recent studies and challenges in mammography that have vast datasets (of over 0.5 million mammograms) have opted for weakly labeled data [14,15].

In general, there are two main approaches to weakly supervised learning, known as image and region based. In image based methods based on CNN [16,17], the input to the model is the whole image. Region inference is then obtained from feature maps after pooling at the final CNN layer (often generating a heat map). In region based methods *e.g.* [9,18], the image is first decomposed into regions. The convolutional layers then process each region separately. Subsequent layers then classify the regions and aggregate results to a global class level.

Image based. Zhu et al. [17] proposed an image-based method for mammogram classification based on Multi-Instance Learning (MIL) that classifies large tiles of the image by max-pooling over feature maps, with sparsity soft constraints. However, when using down-sampled images, their method yielded detection maps with a low resolution of just 6×6 pixels, which curtails practical use considerably. Hwang et al. [16] also took an image-based approach using a CNN with two whole-image classification branches that shared convolution layers. One branch used fully connected layers, and the second branch used 1×1 convolution layers, resulting in one map per class, and then a global max pooling on each map. Their method yielded a low AUROC of 0.65 over 332 MIAS mammograms. Both of these image-level studies [17,16] addressed a binary classification task with a small test set of 410 full-field dig-

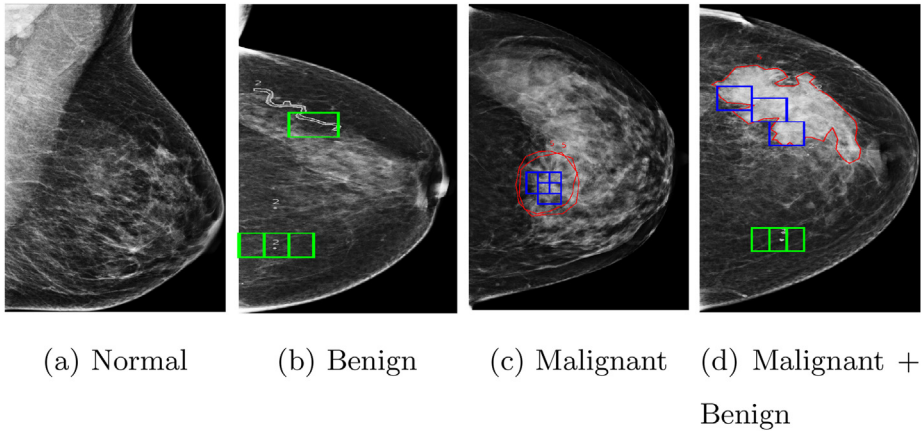


Fig. 1. Illustration of the classification and localization task. The input is the corresponding mammogram and the output is the class (Normal/Benign/Malignant) and the finding localizations. The radiologist's local annotations are shown as contours (red for malignant and gray for benign). The class and the model's local predictions are shown as colored bounding boxes (Normal mammograms do not have a bounding box, Benign regions are in green and Malignant in blue). The image Malignant + Benign is a case with additional benign finding.

ital mammography (FFDM) mammograms [17] or using non-FFDM (digitally scanned) images of the MIAS cohort instead.

Region based. Yan et al. [18] proposed a region-based method for a different use-case of discriminating between local anatomies in CT scans, using MIL in a DNN setting. Choukroun et al. [9] recently implemented a region-based approach with the MIL paradigm to classify the entire mammogram according to the max-probability region, thus also providing detection in *full resolution*.

These methods [9,17,18] apply an implicit detection regime via a max-pooling operation on regions or region classification probabilities.

Dual branch architecture. Recent studies on natural images, suggest that applying explicit *data-driven* detection in parallel to classification yields improved performance [10]. In this study we follow a region-based dual branch approach, but differ and generalize the existing method [10] in two main ways: 1) We don't use any unsupervised region proposal in our scheme as it is commonly unavailable in medical imaging. 2) We adapt the architecture in [10] and extend the *region* classification stream to include an additional *normal-region* class but without any detection counterpart. This makes it possible to handle images without any findings (objects). In addition, it reduces the false positives resulting from normal regions in detection. Also, this enable to use the network in a semi-supervised detection setting with joint learning from weakly and fully supervised data. The extension for handling radiology images assessed as normal is equivalent to images without any objects in natural images. The addition of a normal-region class changes the probability distribution for the regions, and allows improved classification of these specific and prevalent normal cases in many medical use cases such as screening mammography. Similar to [13], we further connect the branches by adding information from the classification branch to guide the detection branch to the most relevant regions.

Our model is capable of multi-class classification and detection that provides localization of the abnormalities in full resolution. We compare our method to the one described in [9] and an approach based on [10]. We report improved performance in both classification and detection.

2.2. Semi supervised detection

Semi supervised detection methods involve the fusion of weak labels with a subset of data having local annotations, namely *fully labeled* (also known as strongly labeled) data. There are two main approaches to semi-supervised detection setting. The first approach is two-stage training with a stage for fully labeled data and a stage for weakly labeled data. The second approach involves joint training from weakly and fully supervised data.

Two-stage approach. A large data set of fully labeled data with lesion annotations is used to train a region based classifier. Then, at a subsequent stage the model is modified for whole image input (usually decomposed into regions) and fine-tuned on the weakly labeled data to create a weakly labeled classifier [5,19,4,20]. However, these methods rely strongly on local annotations and need a sufficiently large fully labeled dataset to initialize the model. They are unable to train solely on weakly labeled mammograms and often lack detection capability (except Ribli et al. [5] that uses detection based on instance labels).

Wu et al. [21] used patch-level classifier for producing heatmaps as additional input channels to a multi-view breast-level classifier.

Joint training from weakly and fully supervised data. A single model is trained jointly using weakly and fully supervised data by combining a weakly supervised objective function with a fully supervised objective function.

Yan et al. [22] proposed a method for weakly supervised training of Fast RCNN [23] via Expectation–Maximization (EM). Focusing on the detection problem, they treated instance-level (region level) labels as missing data for weakly annotated images. Their method alternated between two steps: 1) E-step: estimating a probability distribution over all possible latent locations in weakly supervised images, and 2) M-step: training Fast RCNN using estimated locations from the last E-step. They proposed a semi-supervised learning method by adding a standard fully supervised objective function to the fully supervised images which are then used to train the Fast RCNN network in the M-step in addition to the weakly supervised objective function. Their method was applied on non-medical (natural) images, and in practice, the quality of the solution depended heavily on initialization by another method ([10], which we compare our method with). Furthermore, their approach required thousands of Fast RCNN training iterations at each M-step, which is computationally expensive, particularly for large images such as mammograms.

Cinbis et al. [24] suggested a MIL approach for weakly supervised detection in natural images. They suggested extending their method to a semi-supervised setting by replacing the top region selection, obtained from MIL, with the ground-truth regions when training from fully-supervised images.

In the medical domain, an approach based on Faster RCNN [25] was taken by Shin et al. [26], and was applied to breast Ultrasound (US) images. They also proposed semi-supervised training, but based on combination of Faster RCNN [25] and MIL. However, in breast US, only the field of view with suspicious masses were considered (and not calcifications or images without any abnormality). Unlike mammograms, a lesion in an US captures a relatively large area of the image. Mammography therefore appears to be a greater challenge in that there are more types of lesions with a significantly lower signature.

Li et al. [27] proposed a semi-supervised classification and detection method for chest X-ray images. In their model, the input image is processed by CNN. Then, there is max pooling or interpolation on the feature maps to get patch grid, which is then processed by a fully-convolutional recognition network, resulting in patch scores for multiple categories. Then, they have global predictions based on MIL criterion. They define the global positive probability by the complement of the joint probability of all the patches being negative, assuming patches probabilities are independent of each other. They combine a fully supervised loss function on the fully supervised images and a weakly supervised loss function on the weakly supervised images.

We follow the joint training approach. In our approach, the local annotations are used as auxiliary data, and our model can be trained with a small fully annotated dataset, mostly relying on weak labels. Given the high cost of annotation in many medical domains, we believe that this approach can provide a competitive edge.

3. A dual branch weakly supervised detection methodology for mammograms

In this section we propose a deep network architecture that classifies mammogram regions into three different classes: normal tissue, benign, and malignant findings using labels at the image level (also known as *weak* labels).

We first decompose the image into regions that are fed into the network. The network has two branches: a *classification* branch that computes local probabilities of malignant, benign and normal for each region, and a *detection* branch that ranks regions relative to one another for the malignant class and, independently, for the benign class. The branches are then combined at a subsequent

layer to obtain an image-level decision for the presence of malignant and/or benign findings. The proposed weakly supervised network architecture is depicted in Fig. 2, and the algorithm is summarized in Table 1.

Region extraction. Given a mammography image, we first perform pre-processing to compute feature representations for regions within the breast. To this end, we used a sliding window of 224×224 overlapping regions (with a 112×112 stride) within the breast region excluding the axilla (using a method similar to [28]).

Due to the relatively small training dataset, we employ a two-stage deep neural network architecture. In the first stage, we apply a transfer learning approach by using the pre-trained VGG128 network [29], trained on the ImageNet dataset [30]. In our model, we extract CNN codes from the last hidden layer as 128D feature vectors per region. Then, we process each region separately by a fully connected (FC) layer. Formally, an image x, is first decomposed into m regions denoted by $r_1, \ldots, r_m$ such that $\phi(r_i) \in \mathbb{R}^{128}$ is the feature vector representation of the i-th region.

Classification branch. We first compute a local decision for each region separately. Each region is classified, in this study, as normal (N), benign (B) or malignant (M) using a softmax layer:

$$p_{\text{cls}}(c|r_i) = \frac{\exp(w_c^\top \phi(r_i))}{\sum_{d\in\{N,B,M\}} \exp(w_d^\top \phi(r_i))}, \quad c \in \{N,B,M\}, \quad i = 1,\ldots,m \tag{1}$$

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

Detection branch. In parallel, we compute the relevance of each region for the global image-level decision. We perform a separate detection process for each type of abnormality – one for malignant regions and one for benign regions. The normal class has different characteristics. These regions are prevalent in all types of mammograms, similar to the "background" in natural images. Therefore, the normal class is not associated with a detection scheme (see Fig. 2). This is a novel extension to previous modeling in [10]. In [10] the image-level class set and the region-level class set are the same and are used in both branches. The detection result is a distribution for the malignant class and a distribution for the benign class. Each such distribution is over all the regions in the image implemented by a softmax operation. Formally, let z_c be a hidden random variable representing the localization of class c findings in the image. Then, given an image x, the probability of $z_c = i$ in the c distribution is:

$$p^c_{\text{det}}(i|x) = \frac{\exp(u_c^\top \phi(r_i))}{\sum_{j=1}^m \exp(u_c^\top \phi(r_j))}, \quad c \in \{B,M\}, \quad i = 1,\ldots,m \tag{2}$$

such that u_B and u_M are the parameter-sets of the benign and malignant detectors, respectively. Note that $p^c_{\text{det}}(i|x)$ is equivalent to the ranking of the i-th region in image x relative to the other regions in x for class c.

Image level decision. Given the region-level classification results and the region detection distribution, we can now evaluate the image-level classification. 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 for 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 given mammogram image x are obtained as a weighted average of the local (i.e. region-level) decisions:

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

Comparison to previous dual-branch approach. Since in many medical applications such as mammography, the most prevalent cases are normal without any findings, we extended the method in [10] by adding a normal-region (N) class to the classification branch. Note that in our new scheme the normal class is only added to the classification branch and not to the detection branch or to the image-level class set (see Fig. 2). This is a novel generalization to previous modeling in Bilen et al. [10]. In oppose to Bilen et al. [10], by allowing classification of *regions* to normal, we can handle "clean" images without any findings. Normal images in our model are then discriminated by having a low probability for both M and B findings. The probability for an image to be normal can then be obtained via the joint probability $p(y_M = 0, y_B = 0|x)$.

This extension is also important for reducing the false positives in detection (localization) resulting from normal regions, as shown in Section 5, since normal regions gain high probability for local class N and low probabilities for M and B (instead of expected uniform probabilities over M and B when the N class is not used [10]).

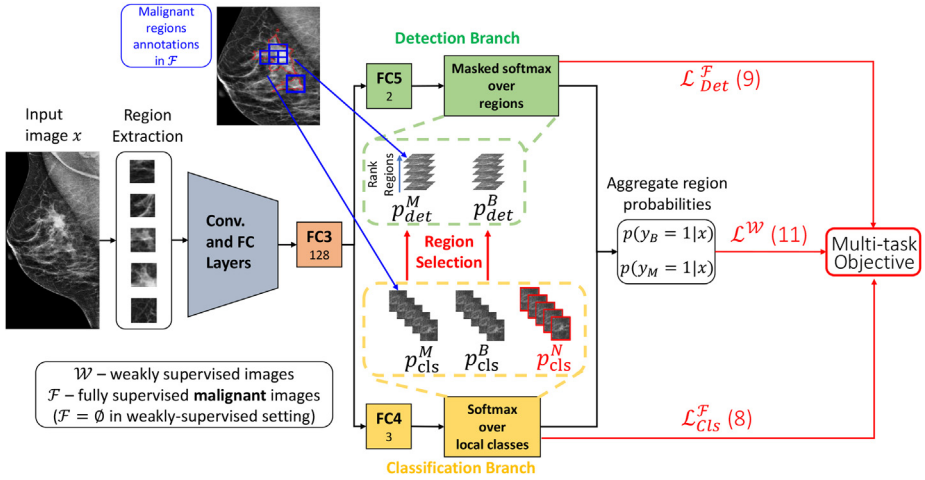


Fig. 2. Architecture overview. The novel elements are in red. Our new scheme has an additional class in the classification branch (p^N_{cls}) with no associated detection and a region selection model connecting the branches. FC blocks show fully-connected layers with the corresponding output size. There are two FC layers in the "Conv. and FC layers" block from the pre-trained VGG. Layer FC3 is shared between branches, FC4 is for the classification branch, and FC5 is for the detection (ranking) branch. $\mathcal{L}^{\mathcal{F}}_{Cls}, \mathcal{L}^{\mathcal{F}}_{Det}$ and $\mathcal{L}^{\mathcal{W}}$ corresponds to Eqs. 8, 9 and 11 respectively.

Table 1
The proposed weakly supervised detection method.

Input data: an image x decomposed into m regions $r_1,\ldots,r_m$ each represented by 128 features computed by a VGG network.
Algorithm:

- Region level classification into classes $C=\{N,B,M\}$:

$$p_{\text{cls}}(c|r_i)=\frac{\exp(w_c^\top\phi(r_i))}{\sum_{d\in C}\exp(w_d^\top\phi(r_i))},\quad c\in C$$

- For $c\in\{B,M\}$:
  - Mask Computation:
    $h_c(i)$ is a binary value indicating whether region i is one of the k regions with the highest probability of being classified as c.
  - Selected regions ranking:

$$p^c_{\text{mask-det}}(i|x)=\frac{h_c(i)\exp(u_c^\top\phi(r_i))}{\sum_{j=1}^m h_c(j)\exp(u_c^\top\phi(r_j))}$$

  - Image level decision:

$$p(y_c=1|x)=\sum_{i=1}^m p^c_{\text{mask-det}}(i|x)p_{\text{cls}}(c|r_i)$$

In addition, this modified architecture enables the use of a fully supervised loss on the classification branch for the extension to a semi-supervised detection setting in Section 4.

Region selection. So far the detection branch's decision has solely been based on the features that were extracted from the image regions. It makes sense to use the classification decision results to guide the detection process. For example, if a region is clearly classified as malignant, it is likely that the malignancy detection will favor this region. Since the classification branch includes an additional class for normal regions, the suspicious regions in the B and M classes can be used to guide the detection branch and create a soft alignment between the branches. We formalize this intuition by a region selection step. Now, let $p_{\text{cls}}(M|r_1),\ldots,p_{\text{cls}}(M|r_m)$ be the region probabilities of being classified as malignant. In the malignant detection process, we only consider the k regions with the highest probability of being classified as malignant and only apply the softmax operation on these selected regions. Let $h_M(i)$ be a binary value indicating whether region i has been selected for the malignancy detection process. We can apply the same selection criterion to the benign detector. Thus, each detector's ranking is conducted solely on the relevant regions according to the classification branch. In the modified detection branch we replace the softmax over regions by a masked softmax:

$$p^c_{\text{mask-det}}(i|x)=\frac{h_c(i)\exp(u_c^\top\phi(r_i))}{\sum_{j=1}^m h_c(j)\exp(u_c^\top\phi(r_j))},c\in\{B,M\},\quad i=1,\ldots,m. \tag{4}$$

This paradigm guides the M detector to focus on the most probable malignant regions in malignant mammograms. However, if the image is normal or contains a benign finding, the model will concentrate on regions that were most probably and erroneously classified as malignant (hard negatives). This process, which is applied similarly to the benign class is equivalent to *hard negative mining*. In the experimental section we compare network architectures with and without masked detectors and show that applying region selection yields superior performance.

Training. Assume we are given a set of n weakly labeled mammography images $\{x(1),\ldots,x(n)\}$. Each image $x(t)$ consists of regions $\{r_1(t),\ldots,r_m(t)\}$ and is associated with a binary tuple label $(y_M(t),y_B(t))$ that indicates whether the image contains at least one malignant and/or one benign finding respectively. A normal case will have a $(0,0)$ label whereas a mammogram with both M and B finding will be labeled $(1,1)$. The network provides soft decisions for each image $x(t)$ regarding the values of $y_M(t)$ and $y_B(t)$. The objective function that we maximize in the network training step is the following likelihood function:

$$L(\theta)=\sum_{c\in\{M,B\}}\sum_{t=1}^n\log p(y_c(t)|x(t);\theta) \tag{5}$$

such that θ is the parameter-set of the model (which includes the fully connected layer ϕ and the parameters w and u) and the probability $p(y_c(t)|x(t);\theta)$ is defined in Eq. (3).

4. Semi supervised detection methodology

4.1. Approach overview

In this section, we extend our weakly supervised setting to a novel semi-supervised approach. In a semi-supervised setting, we assume that part of the weakly labeled data has been subjected to local annotations, thus generating a subset of *fully-labeled* data. This local annotation can take the form of contours around lesions or simply bounding boxes. We demonstrate our model on M vs. B $\cup$ N. To reduce the annotation workload, let us assume that the malignant class has a fully-labeled subset in which only the malignant findings are locally annotated (note that malignant images can still include benign findings).

We make use of different ratios of local annotations in the malignant class (25–100%) to present the impact of these annotations on performance. Due to the rarity of malignant findings with respect to benign ones, the annotated set only captures 2.5–10% of all the lesions in the cohort, therefore demanding a low workload for annotation.

Our dual-branch approach differs from previous approaches [22,26,27] in architecture and objective function. Our semi-supervised method is different from previous methods by having a region ranking branch in the architecture. In addition, previous methods [22,26,27] added a fully-supervised objective function on the region classification in fully-supervised images subset. In our method, we add a fully-supervised objective function on the region classification, and, in addition, we add a fully-supervised objective function on the detection (ranking) branch's region probabilities of the fully-supervised images subset.

4.2. Semi-supervised detection objective function

Although local annotations on a large scale are commonly out of reach [15], in this section we examine the effect of engaging with a small set of locally annotated data combined with a large set of

weakly labeled data. We assume that the training set contains two distinct sets, one with weakly and one with fully labeled images. We denote $\mathcal{W}$ as the set of indices of the weakly-labeled images (these can be malignant, benign or normal) and $\mathcal{F}$ as the set of indices of the *fully-labeled* images; namely, mammograms where lesions have been locally annotated. For each fully labeled image, x, we are given a set $\mathcal{M}_x$ of malignant regions. We next describe how we transform the pixel-level information (i.e. contour annotations) into the region-level labels based on the intersection between our extracted regions and the malignant lesion. To this end, we define a soft version of Intersection over Union (*IoU*) called the *Intersection over Minimum* (*IoM*). This measure computes the ratio between the area of the intersection with respect to the minimum size between the i-th region r_i and the lesion area:

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

where c is the annotated domain. In our setting the region size is fixed and the lesion scale can vary by a factor of 10. This definition therefore allows a positive region to cover a small lesion or alternatively be located within a large finding. We define the local label of a region as malignant (M) if the region has $IoM \geqslant \alpha$ with a ground-truth (GT) **malignant** finding, and define the label as either benign or normal (BN) if the region has an empty intersection with all the GT malignant findings. We set $\alpha = 0.5$. Formally, the label of region r_i, denoted by y_i, is defined as follows:

$$y_i = \begin{cases} M & \exists\ c \in \mathcal{M}_x \ \ s.t.\ IoM(i,c) \geqslant \alpha \\ \\ BN & \forall\ c \in \mathcal{M}_x,\ r_i \cap c = \varnothing \end{cases} \tag{7}$$

Non-malignant regions with $IOM < \alpha$ are ignored during training. In practice, we achieved better performance when ignoring those regions during training compared to labeling the regions as BN.

In order to engage the local annotations, we propose two separate and novel objective functions that are imposed directly on the region classification and detection probabilities. In the fully supervised objective of the classification branch, we compute the log likelihood according to the region true classes (as M or BN) as:

$$\mathcal{L}^{\mathcal{F}}_{Cls}(\theta) = \sum_{t\in\mathcal{F}}\sum_{i} \log p_{\text{cls}}(y_i(t)|r_i(t)) \tag{8}$$

where t goes over all the fully labeled images, and i goes over the labeled regions in each image. The probability of a region to be classified as malignant, $p_{\text{cls}}(M|r_i(t))$, is defined in Eq. (1), and $p_{\text{cls}}(BN|r_i(t))$ is the complement probability (i.e., the probability of being classified as either benign or normal).

In the fully-supervised objective of the detection branch, we want to concentrate on the malignant regions. We therefore define the detection branch objective as:

$$\mathcal{L}^{\mathcal{F}}_{Det}(\theta) = \frac{1}{|\mathcal{F}|}\sum_{t\in\mathcal{F}} \log\left(\sum_{i|y_i(t)=M} p^{M}_{det}(i|x(t))\right). \tag{9}$$

This demanding regions with high overlap over M-lesions to have high M-probability. This soft constraint alters the weakly supervised decisions toward manually labeled regions. The trained model eventually relies on discriminative power and similarity to the annotated regions as the source of malignancy when making its decisions.

Without loss of generality, we assume the fully-supervised objective is applied on the malignant images in $\{x(t) : t \in \mathcal{F}\}$. Our final fully supervised objective is then obtained as:

$$\mathcal{L}^{\mathcal{F}}(\theta) = \lambda_1 \mathcal{L}^{\mathcal{F}}_{Cls}(\theta) + \mathcal{L}^{\mathcal{F}}_{Det}(\theta). \tag{10}$$

We set $\lambda_1 = \beta/m_f$ where m_f is the total number of regions in the train data that have a region-level label. For simplicity, we set $\beta = 1$.

The weakly supervised part, $\mathcal{L}^{\mathcal{W}}$, is defined in a similar way as in Section 3, Eq. (5). In the semi-supervised setting, this objective is defined over the weakly labeled training subset for the M class and over all the images for the B class:

$$\mathcal{L}^{\mathcal{W}}(\theta) = \frac{1}{|\mathcal{W}|}\sum_{t\in\mathcal{W}} \log p(y_M(t)|x(t);\theta) + \frac{1}{n}\sum_{t=1}^{n} \log p(y_B(t)|x(t);\theta) \tag{11}$$

In order to prevent redundancy in the training samples we avoid using the fully labeled images also as weakly labeled samples, since they were shown to degrade performance in Shin et al. [26].

The fusion of the weakly and fully supervised settings can now be achieved by maximizing the following multi-task objective:

$$\mathcal{L}(\theta) = \mathcal{L}^{\mathcal{W}}(\theta) + \lambda_2 \mathcal{L}^{\mathcal{F}}(\theta) \tag{12}$$

where $\mathcal{L}^{\mathcal{W}}$ denotes the *weakly* supervised part and $\mathcal{L}^{\mathcal{F}}$ denotes the *fully* supervised part.

5. Experimental results

5.1. Experiment setup

Dataset. We conducted experiments on a large screening dataset, named IMG, with full field digital mammography (FFDM). The cohort was acquired from different Hologic devices and 4 different medical centers (with approximately 3 K × 1.5 K 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, pathologies (including benign and malignant cases) and breast densities that corresponds to what is typically found in screening clinics. The dataset was composed of 2,967 mammograms with normal images as well as various benign and suspiciously malignant findings. In terms of the *global* image BI-RADS (Breast Imaging Reporting and Data System), we had 350, 2,364, 146 and 107 corresponding to BI-RADS 1,2,4 and 5 captured from 65, 693, 81 and 62 individuals respectively. Note that our BI-RADS 1 (Normal category) did not contain any suspicious findings, or confidently benign ones. Since a mammogram can contain findings with different BI-RADS categories, the global image BI-RADS was set by the most severe finding in the image (max operation), and the global patient BI-RADS was set by the max global image BI-RADS for that patient in a specific study, according to clinical guidelines.

Mammograms with global BI-RADS of 3 were excluded from our IMG dataset since these intermediate BI-RADS are commonly assigned based on other modalities (*e.g.* ultrasound) and comparison to prior mammograms [31] which are often unavailable. However, our data set included BI-RADS 3 findings that were not the most severe ones in the image. In terms of breast composition, 20% were "almost entirely fatty", 48% had a "scattered fibroglandular density", 27% were "heterogeneously dense" and 5% were "extremely dense". With respect to the dominant pathologies, our data set included 4525 calcifications (micro and macro) and 926 masses.

In our test scenario, we split the mammograms into the following three global labels: BI-RADS 4 & 5 were defined as malignant (M), BI-RADS 2 were defined as benign (B) and BI-RADS 1 as normal (N). We included all types of suspiciously malignant abnormalities in the M class such as mass, calcification, architectural distortions etc. This discrimination in data classes creates a specific challenge,

demanding the model to distinguish between images with very similar types of lesions, such as malignant versus benign masses or different types of micro-calcifications that are often ambiguous even for expert radiologists. BI-RADS-based class separation is frequently used (e.g, [15,19,32–36]) often because of the lack of pathological results in the dataset and the need to construct a large positive set. In Shen et al. [19], the authors claimed that although the INbreast dataset includes pathology results, they use BI-RADS assessments for class labels, due to "lack of reliable pathological confirmation". In a similar way, they defined all images with BI-RADS 1 and 2 as negative and BI-RADS 4, 5 and 6 as positive.

Our second test bed used for our weakly supervised model, was composed of the INbreast (INB) publicly available FFDM dataset [8]. This small dataset has 410 mammograms from 116 cases and was split into 100 positive (global BI-RADS 4,5,6) and 310 negative (global BI-RADS 1,2,3) mammograms. Note that in this case we included BI-RADS 3 to enable comparison with previous methods in literature. We conducted a random patients split on the INbreast images with 50% for train and 50% for test.

Implementation. We implemented our model in the TensorFlow framework using the Adam optimizer for training, with a learning rate of 10^{-4}, dropout of 0.5, l_2-regularization and a batch-size of 256 images. This included all the regions from each image (on average approx. 200). We initialized the weights of the shared fully connected (FC) layer with a normal distribution [37]. The weights of the FC layers in the branches were initialized with zero mean and 10^{-4} STD normal distribution. For the number of selected regions we chose $k = 10$ (other values were tested but yielded lower performance). We set $\lambda_2 = 1$ in our semi-supervisded experiments, and we discuss other values in Section 5.3. To enlarge and balance the training set, we used augmentations by adding rotations of $7 \times 45^{\circ}$, left–right and up–down flips and 6 image shifts.

Evaluation Procedure. Our evaluation on IMG dataset was based on 5-fold patient-wise cross-validation, where at each train and test iteration, all the images from the patient under test were strictly excluded from the training set. To this end we randomly split the dataset into 5 folds according to patient IDs, maintaining a similar distribution over breast composition and lesion types in the folds. All the performance values were based on the average over random split, 5-fold cross validation.

Compared Models. As our model outputs two probabilities per image ($p(y_c = 1|x)$, Eq. (3)), we can create 2D probabilities maps and conduct multi-class classification. However, to compare our results to previous methods and as an instance of a practical use case, we evaluated system performance on two binary classification tasks by joining two "nearby" classes; namely, M with B or B with N. 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). For performance measures, in addition to AUROC, we also report two other practical measures as used in [14]. The partial-AUC ratio (pAUCR) associated with the ratio of the area under the ROC curve in a high sensitivity range ([0.8,1]) represents the AUROC in a more relevant domain for clinicians. In addition, we report the specificity extracted from the ROC curve at sensitivities of 0.85 and 0.90 that represent an average operation point (OP) for expert radiologists, as reported in [38].

5.2. Classification results

We compare of our model's performance to several baselines. We then discuss the impact of the fully labeled data engaged with our multi-task loss. For evaluation, we present our results on the two binary classification tasks, M vs. BN and MB vs. N. In addition to the proposed Cls-Det-RS model, we implemented three baselines, 1) *Max-Region* [9] presenting a region classification only approach with max over regions, 2) the *DB-Baseline* presenting a dual-branch approach equivalent to Bilen et al. [10] and 3) the *Cls-Det* as our approach without region selection.

Weakly-supervised setup. Table 2 presents performance for the two binary classification tasks. Considering purely the weakly labeled dataset, our method (Cls-Det-RS) outperformed the DB-Baseline and Max-region [9] on all measures and in both classification scenarios. The results of the model without the region selection (RS) showed that in average, the addition of region-selection indeed improved performance. We further conducted a breast level analysis by considering both views of the same breast. To this end we assigned the max probability between the views to the specific breast. The results exhibited similar performance to the single mammogram processing.

Train and test on the small public data set of INB yielded AUROC of 0.73. Note that this result is without using an external **fully labeled** data set in oppose to [4,5]. This result shows the performance of our model when trained on a very small data set. It is further comparable to AUROC 0.74, reported in [33] when trained on single MG, yet used fully supervised data.

Semi-supervised setup. Next we analyze the performance of our semi-supervised model. In order to reduce the demand for local annotations, we only considered local annotations for the malignant findings in our setting. We opted for the classification task of M vs. BN as commonly considered in previous works [16,17,33]. We further evaluated the impact of the ratio of the fully supervised train set as a measure of the cost effectiveness of the annotation workload. The results for our semi-supervised setting (Cls-Det-RS) are shown in Table 2. The classification performance improved as more localized regions are used. This continued up to 100% utilization of the local annotation (fully supervised).

5.3. Detection results

Although the train process begins without any labels on regions, the impact of each region can be scored after the training process by:

$$d^c(r_i) = p_{cls}(c|r_i)p^c_{det}(i|x), \quad c \in \{B, M\}, \quad i \in \{1, \ldots, m\} \tag{13}$$

The top k regions for each class (B/M) can now be visualized and compared to the radiologist's annotations as the source of malignancy or benign class of the image. Fig. 3 shows several examples with localization in the test set, overlaid with the radiologist's annotations (used only for validation). As observed, the method is capable of separately highlighting multiple types of abnormalities such as benign and malignant lesions without having an instance level annotation.

We further evaluated our localization performance by a quantitative measure. Targeting the localization as the system's self-explanation tool, we used a less strict measure than the standard intersection over union (IoU) for correctness of our localization outcome. We follow the weak localization as intersection over the *minimum* area between the region and the lesion (IoM) as defined in Eq. (6) (also used in [39]). This measure allows explanation of an outcome when a specified region contains a true type of lesion or vice versa. Since our region size is relatively small and fixed, this setup will not allow over-sizing of the localization area (see examples in Fig. 3). Unlike previous methods of [16,17] we formally asses the accuracy of our localization results by Eq. (13).

For an image classified as c, we consider all the regions with $d^c(r_i)$ over a certain threshold. Correct localization per lesion is obtained if $IoM \geqslant 0.5$. We present the free-response receiver operating characteristic (FROC) localization accuracy for class $c \in \{M, B\}$ using $d^c(r_i) \geqslant Threshold$. The detection sensitivity in

Table 2
Binary classification performance compared to previous methods in weakly, semi and full supervised settings.

Method	AUROC	pAUCR	Spec @ Sens	
			0.85	0.90
M vs. BN: Weakly-supervised methods				
DB-Baseline [10]	0.709 ± 0.020	0.251 ± 0.05	0.37	0.27
Max-Region [9]	0.699 ± 0.047	0.235 ± 0.10	0.36	0.24
Cls-Det	0.710 ± 0.026	**0.280** ± 0.06	**0.42**	**0.31**
Cls-Det-RS	**0.728** ± 0.036	0.275 ± 0.10	0.40	0.27
MB vs. N: Weakly-supervised methods				
DB-Baseline [10]	0.826 ± 0.01	0.347 ± 0.03	0.51	0.37
Max-Region [9]	0.817 ± 0.02	0.323 ± 0.07	0.48	0.35
Cls-Det	0.832 ± 0.02	0.355 ± 0.06	0.51	0.36
Cls-Det-RS	**0.841** ± 0.02	**0.367** ± 0.05	**0.55**	**0.38**
M vs. BN: Semi-supervised methods				
SS Cls-Det-RS.25	0.731 ± 0.029	0.305 ± 0.108	0.40	0.31
SS-Baseline-RS.5	0.740 ± 0.022	0.316 ± 0.126	0.43	0.30
SS Cls-Det-RS.50	**0.745** ± **0.032**	0.313 ± 0.119	**0.46**	**0.33**
SS Cls-Det-RS.75	**0.745** ± **0.026**	**0.320** ± **0.109**	0.42	**0.33**
M vs. BN: Fully-supervised (on M class) method				
SS Cls-Det-RS 1.0	0.751 ± 0.026	0.316 ± 0.078	0.47	0.32

The bold is used to indicate the best result in the column.

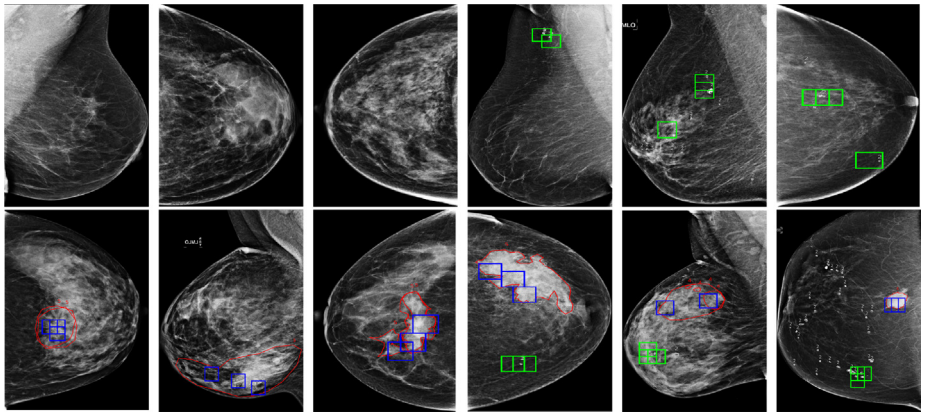


Fig. 3. Localization success in the weakly labeled setting. True malignant and benign lesions are annotated in red and gray respectively. Top 3 – M (blue) and B (green) regions are shown. Note the correlation between the radiologist's annotation and the model's predictions for each class. The top 3 images on the left are normal images without findings, where no bounding boxes were predicted. The 3 right hand images in the top row show cases of only benign findings. The lower 3 left hand images only have malignant findings, and the 3 right hand images have both malignant and benign findings. Note the agreement between the ground truth location and class of the finding with our predictions, without having any instance annotations in the training set. Best viewed in color.

the FROC is the fraction of images in the True-Positive set with at least one correct localization. The results show that the region selection yielded the best performance with relatively low False positive per image (FPPI).

Weakly-supervised setup. Fig. 4 shows the detection performance as FROC. Performance for MB vs. N is shown on the left. Although at low FPPI, DB-Baseline (dotted black curve) and our model (Cls-Det-RS, dashed orange curve) are comparable, at high detection sensitivity our model shows slightly improved performance. However, our model clearly outperforms Max-Region [9] (dotted red curve).

Fig. 4 right plot depicts FROC curves for detection of the malignant lesions (BI-RADS 4 & 5). In this set-up, we first compare our weakly supervised model to several baselines and then show the impact of our semi-supervised network with various ratios of fully labeled data. In particular, the detection performance in our weakly supervised model (dashed orange) is compared with the DB-Baseline [10] (dotted black) and the Max-Region method [9] (dotted red). In this scenario of detecting malignant lesions, the DB-Baseline shows poor results. Although the Max-Region shows improvement over DB-Baseline, our model clearly outperforms both. In addition, our model with region selection (Cls-Det-RS, dashed orange), outperforms our model without region selection (Cls-Det, dashed blue).

Semi-supervised setup. The right plot in Fig. 4 shows that including local annotations in our semi-supervised model (SS-Cls-Det-RS) improves detection. The green lines indicate results when using different ratios of fully labeled data (wider curves indicate higher fully labeled ratio in training). The wide cyan line stands for full supervision on the **M class**. The detection sensitivity further improved when more locally annotated mammograms were used. However, the influence of local annotations plateaus approaching the 75% ratio (SS-Cls-Det-RS 0.75), presenting similar performance to the fully-supervised method (SS-Cls-Det-RS 1). The performance drop in M vs. BN compared to MB vs. N (right vs. left plot in Fig. 4), indicates the model's difficulty in distinguishing between benign and malignant lesions, as often is the case with radiologists.

Setting the value of λ_2. The parameter λ_2 controls the balancing between the fully supervised images and the weakly supervised images (that contain both benign and malignant images). We found that the classification and localization result are insensitive

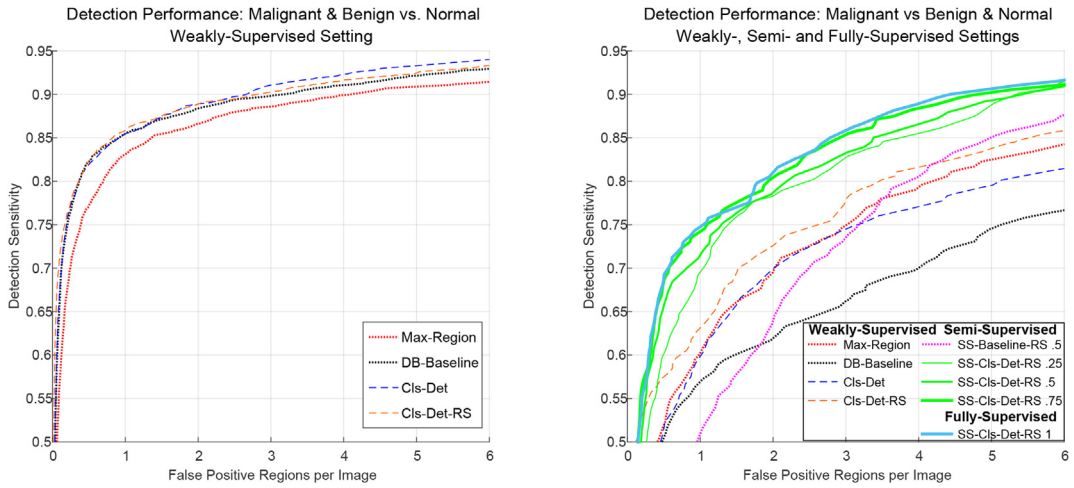


Fig. 4. FROC for detection performance at an operation point of 0.85 sensitivity in classification. Left: MB vs. N for the weakly-supervised setting. Right: M vs. BN, comparing weakly-, semi- and fully- supervised settings. Baselines: Max-region [9], DB-Baseline[10]. SS-Baseline-RS.5: our semi-supervised approach with 50% fully-supervised data when the fully-supervised objective is only on the classification branch. Weakly supervised proposed methods with and without region selection: Cls-Det, Cls-Det-RS. Proposed semi-supervised methods: SS-Cls-Det-RS with various ratios of fully supervised data (indicated by green line with increasing width as a function of the fully-supervised data ratio). Fully-supervised method: SS-Cls-Det-RS 1. Best viewed in color.

to the values of λ_2 in a range of roughly $0.1-1$. Increasing λ_2 yields lower performance in the benign class without improving the classification performance of the malignant class and with slight improvement in the localization of the malignant class (e.g. for $\lambda_2 = 5$, we get 0.72 MB vs. N AUC, and for $\lambda_2 = 10$, this decreases to 0.65 AUC). As decreasing λ_2, we get a slightly worse classification performance of M vs. BN and worse performance in the localization. Fig. 5 shows FROC of our semi-supervised approach, SS-Cls-Det-RS, with 50% fully-supervised data for various values of λ_2.

The impact of loss on the detection branch. To this end, we ran our model with loss solely on the classification branch (similar to [22]). We trained our model with 50% fully labeled data, without the detection loss in Eq. (10) (setting $\mathcal{L}^{\mathcal{F}}_{Det}(\theta) = 0$). The resulting FROC (SS-Baseline-RS.5 – dotted pink curve) appears in Fig. 4-right. Comparison to our model (SS-Cls-Det-RS.5 – green) indicates a significant drop of FROC in this baseline, and points to the contribution of our novel detection loss.

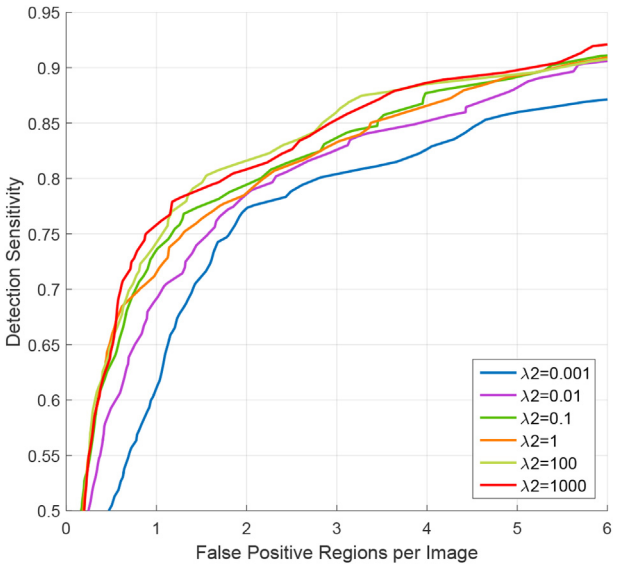


Fig. 5. FROC of M vs. BN detection performance at an operation point of 0.85 sensitivity in classification for several values of λ_2.

5.4. Further analysis on classification results

In this subsection, we presents a visual analysis of the classification results of our weakly supervised method, and multi-class classification results.

Using our multi-label probability output we plot each sample in a probability plane representing the global prediction results of the images. In this plane, each image is a $2D$ point with coordinates as $p(y_M = 1|x)$ and $p(y_B = 1|x)$ probabilities. Fig. 6 shows the global probability plane on a train and test set color coded by the true class. Blue normal (N) images (without any finding) are mostly located near the origin, with low $p(y_M = 1|x)$ and $p(y_B = 1|x)$ showing approximately zero probabilities for malignancy and benign. Green represents images with only benign findings (B). Those are likely concentrated around (0,1) with low $p(y_M = 1|x)$ and high $p(y_B = 1|x)$. Red points represent malignant images without benign findings. Those are emerged at the right side of the plot with high $p(y_M = 1|x)$ and low $p(y_B = 1|x)$. Finally, black points, representing malignant mammograms that also include benign findings, are more likely located in the top-right corner with high $p(y_M = 1|x)$ and high $p(y_B = 1|x)$.

6. Conclusion

In this work, we proposed a method for multi-class classification of mammograms and detection of abnormalities in weakly and semi-supervised settings. We addressed the problem of fusion between weak labels and local annotations in the dataset via a novel objective function. As local annotations are prohibitively expensive in the medical domain, our semi-supervised approach allows reaching nearly fully labeled data performance with a fraction of local annotations. The new model relies mainly on weakly labeled data and therefore can run without any local annotations is the dataset.

We demonstrate our method on a large dataset, and compare our approach with various measures, to several baselines and as well as direct comparison to a previously published method. The results show improvement in AUROC, with a significant performance boost in partial AUC and a practical operation point. Locally annotating only 5% of the data yielded a 10% increase in specificity (at 0.85 sensitivity) that is estimated to lead to yearly 3.6 million

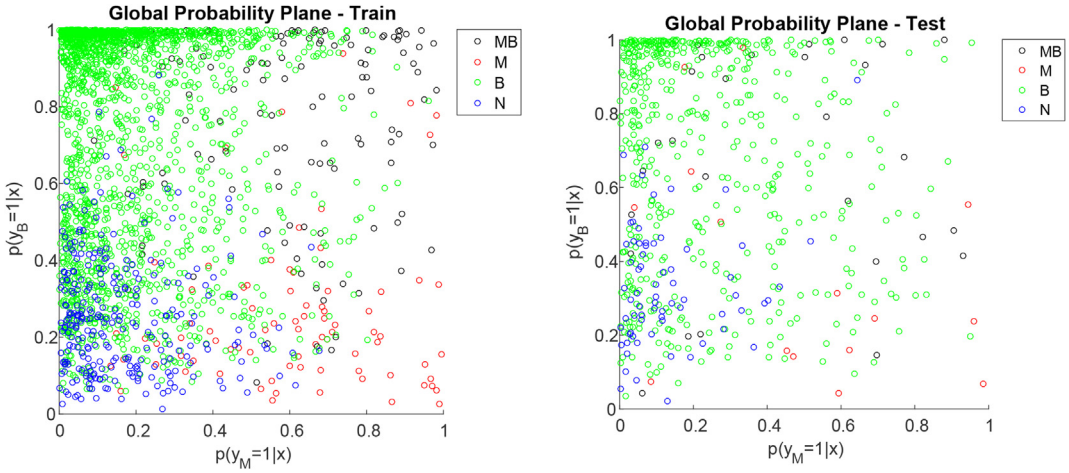


Fig. 6. Global probabilities plane for the train and test set. The figure shows the probability of at least one malignant or benign finding appearing in the mammogram. The samples are color coded to depict prediction accuracy, N-Normal, B-Benign, M-Malignant and MB-Benign and Malignant findings present in the mammogram. Best viewed in color.

fewer false positives in screening mammography [38]. Our method can learn solely from image-level labels, and utilize possibly existing local annotations as bounding boxes around lesions.

A major feature of our system is the localization of the image level decision. This makes system decision interpretable to physicians who obtain the automatic decision. We evaluated our localization performance *quantitatively*, in *full resolution*. The results compete favorably with a previous weakly supervised method and significantly improve in our semi-supervised approach. In the era of Big Data, the combination of large weakly labeled data sets with partially local annotations can provide a cost-effective solution for future decision support systems in medical imaging.

Possible applications suggest second reader in screening mammography and other imaging domains. System explanation based on lesion localization and category should encourage trust among radiologists and is necessary in cases where a quick over-rule is needed if the system decision was found to be wrong.

Our method was evaluated, based on BI-RADS assessment by radiologists. We opted for this setting in order to have a large dataset of approximately 3 K mammograms, as pathologies were not available for all of our high BI-RADS exams. BI-RADS 4 and 5 have positive predictive values of approx. 35% and over 95% respectively and are particularly rare in the population. There are several recent works trained and tested on large FFDM mammogram datasets with pathologies such as [4,5] which used the DREAM Challenge dataset, or [40]. Unfortunately, these datasets are not publicly available and cannot be used by other researchers for benchmarking. We believe that our scenario based on BI-RADS assessments can provide a valid platform for comparison between different methods and baselines. We tested our method and all compared methods on the same data setting to allow for fair comparison.

Our method was limited to analyzing each view separately, without bilateral breast comparison as conducted by radiologists. We intend to use this additional information in our future work to extract correlations between image views and dissimilarities between breast sides.

Combining the proposed approach with end-to-end training of the backbone network is applicable with larger datasets. End to end training as well as using multiple scale and aspect ratio regions constitute interesting future research directions that are beyond the scope of this work.

CRediT authorship contribution statement

Ran Bakalo: Methodology, Software, Writing - original draft, Writing - review & editing. **Jacob Goldberger:** Methodology, Software, Writing - original draft, Writing - review & editing. **Rami Ben-Ari:** Methodology, Software, Writing - original draft, Writing - review & editing.

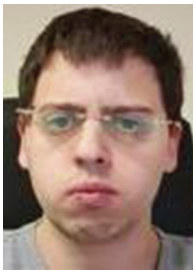
Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] A. Jemal, F. Bray, M.M. Center, J. Ferlay, E. Ward, D. Forman, Global cancer statistics, CA: A Cancer Journal for Clinicians 61 (2) (2011) 69–90.
[2] F.S.S. de Oliveira, A.O. de Carvalho Filho, A.C. Silva, A.C. de Paiva, M. Gattass, 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.
[3] C. Jen, S. Yu, Automatic detection of abnormal mammograms in mammographic images, Expert Systems with Applications 42 (6) (2015) 3048–3055.
[4] W. Lotter, G. Sorensen, D. Cox, A multi-scale CNN and curriculum learning strategy for mammogram classification, in: Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2017.
[5] D. Ribli, A. Horváth, Z. Unger, P. Pollner, I. Csabai, Detecting and classifying lesions in mammograms with deep learning, Scientific Reports 8 (1) (2018) 4165.
[6] A. Katalinic, C. Bartel, H. Raspe, I. Schreer, Beyond mammography screening: quality assurance in breast cancer diagnosis (the QuaMaDi project), Breast Journal Cancer 96 (1) (2007) 157–161.
[7] M.Y. Guan, V. Gulshan, A.M. Dai, G.E. Hinton, Who said what: Modeling individual labelers improves classification, in: American Association for Artificial Intelligence (AAAI), 2018.
[8] I.C. Moreira, I. Amaral, I. Domingues, A. Cardoso, M.J. Cardoso, J.S. Cardoso, Inbreast: toward a full-field digital mammographic database, Academic Radiology 19 (2) (2012) 236–248.
[9] Y. Choukroun, R. Bakalo, R. Ben-Ari, A. Akselrod-Ballin, E. Barkan, P. 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, The Eurographics Association, 2017.
[10] H. Bilen, A. Vedaldi, Weakly supervised deep detection networks, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2846–2854.
[11] R. Bakalo, R. Ben-Ari, J. Goldberger, Classification and detection in mammograms with weak supervision via dual branch deep neural net, in: IEEE International Symposium on Biomedical Imaging (ISBI), 2019.

[12] M. Oquab, L. Bottou, I. Laptev, J. Sivic, Is object localization for free? - weakly-supervised learning with convolutional neural networks, in: Computer Vision and Pattern Recognition (CVPR), 2015.
[13] W. Jiang, T. Ngo, B. S. Manjunath, Z. Zhao, F. Su, Optimizing region selection for weakly supervised object detection, CoRR abs/1708.01723 (2017).
[14] DREAM, The Digital Mammography DREAM Challengehttps://www.synapse.org/#!Synapse:syn4224222.
[15] K. J. Geras, S. Wolfson, Y. Shen, S. G. Kim, L. Moy, K. Cho, High-resolution breast cancer screening with multi-view deep convolutional neural networks, in: arXiv:1703.07047, 2017.
[16] S. Hwang, H. Kim, Self-transfer learning for weakly supervised lesion localization, in: Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2016, pp. 239–246.
[17] W. Zhu, Q. Lou, Y. S. Vang, X. Xie, Deep multi-instance networks with sparse label assignment for whole mammogram classification, in: Medical Image Computing and Computer Assisted Intervention (MICCAI), 2017, pp. 603–611.
[18] Z. Yan, Y. Zhan, Z. Peng, S. Liao, Y. Shinagawa, S. Zhang, D.N. Metaxas, X.S. Zhou, Multi-instance deep learning: Discover discriminative local anatomies for bodypart recognition, IEEE Transactions on Medical Imaging 35 (5) (2016) 1332–1343.
[19] L. Shen, End-to-end training for whole image breast cancer diagnosis using an all convolutional design, in: NIPS Workshop on Machine Learning for Health, 2017.
[20] L. Shen, L.R. Margolies, J.H. Rothstein, E. Fluder, R. McBride, W. Sieh, Deep learning to improve breast cancer detection on screening mammography, Scientific Reports 9 (1) (2019) 1–12.
[21] N. Wu, J. Phang, J. Park, Y. Shen, Z. Huang, M. Zorin, S. Jastrzkebski, T. Févry, J. Katsnelson, E. Kim, S. Wolfson, U. Parikh, S. Gaddam, L. L. Y. Lin, K. Ho, J. D. Weinstein, B. Reig, Y. Gao, H. Toth, K. Pysarenko, A. Lewin, J. Lee, K. Airola, E. Mema, S. Chung, E. Hwang, N. Samreen, S. G. Kim, L. Heacock, L. Moy, K. Cho, K. J. Geras, Deep neural networks improve radiologists' performance in breast cancer screening, in: arXiv preprint arXiv:1903.08297, 2019.
[22] Z. Yan, J. Liang, W. Pan, J. Li, C. Zhang, Weakly- and semi-supervised object detection with Expectation-Maximization algorithm, in: arXiv:1702.08740, 2017.
[23] R. B. Girshick, Fast R-CNN, in: Int. Conference on Computer Vison (ICCV), 2015, pp. 1440–1448.
[24] R.G. Cinbis, J.J. Verbeek, C. Schmid, Weakly supervised object localization with multi-fold multiple instance learning, IEEE Transactions on Pattern Analysis and Machine Intelligence 39 (1) (2017) 189–203.
[25] S. Ren, K. He, R.B. Girshick, J. Sun, Faster R-CNN: towards real-time object detection with region proposal networks, IEEE Transactions on Pattern Analysis and Machine Intelligence 39 (6) (2017) 1137–1149.
[26] S.Y. Shin, S. Lee, I.D. Yun, S.M. Kim, K.M. Lee, Joint weakly and semi-supervised deep learning for localization and classification of masses in breast ultrasound images, IEEE Transactions on Medical Imaging 38 (3) (2019) 762–774.
[27] Z. Li, C. Wang, M. Han, Y. Xue, W. Wei, L. Li, L. Fei-Fei, Thoracic disease identification and localization with limited supervision, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.
[28] C. Chen, G. Liu, J. Wang, G. Sudlow, Shape-based automatic detection of pectoral muscle boundary in mammograms, Journal of Medical and Biological Engineering 35 (2015) 315–322.
[29] K. Chatfield, K. Simonyan, A. Vedaldi, A. Zisserman, Return of the devil in the details: Delving deep into convolutional nets, in: British Machine Vision Conference (BMVC), 2014.
[30] J. Deng, W. Dong, R. Socher, L. Li, K. Li, F. Li, Imagenet: A large-scale hierarchical image database, in: Computer Vision and Pattern Recognition (CVPR), 2009, pp. 248–255.
[31] A.Y. Michaels, R.L. Birdwell, C.S. Chung, E.P. Frost, C.S. Giess, Assessment and management of challenging BI-RADS category 3 mammographic lesions, RadioGraphics 36 (5) (2016) 1261–1272.
[32] N. Dhungel, G. Carneiro, A. P. Bradley, The automated learning of deep features for breast mass classification from mammograms, in: Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2016.
[33] N. Dhungle, G. Carnerio, A. P. Bradley, Fully automated classification of mammograms using deep residual neural networks, in: IEEE Int. Symposium on Biomedical Imaging (ISBI), 2017.
[34] W. Zhu, Q. Lou, Y. S. Vang, X. Xie, Deep multi-instance networks with sparse label assignment for whole mammogram classification, in: Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2017.
[35] A. Akselrod-Ballin, L. Karlinsky, S. Alpert, S. Hasoul, R. Ben-Ari, E. Barkan, A region based convolutional network for tumor detection and classification in breast mammography, in: MICCAI Workshop on Deep Learning and Data Labeling for Medical Applications, 2016.
[36] A. Akselrod-Ballin, L. Karlinsky, A. Hazan, R. Bakalo, A. B. Horesh, Y. Shoshan, E. Barkan, Deep learning for automatic detection of abnormal findings in breast mammography, in: MICCAI Workshop on Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, 2017.
[37] X. Glorot, Y. Bengio, Understanding the difficulty of training deep feedforward neural networks, in: Int. Conference on Artificial Intelligence and Statistics (AISTATS), 2010, pp. 249–256.
[38] C.D. Lehman, R.F. Arao, B.L. Sprague, janie M. Lee, D.S. Buist, K. Kerlikowske, L. M. Henderson, T. Onega, A.N.A. Tosteson, G.H. Rauscher, D.L. Miglioretti, National performance benchmarks for modern screening digital mammography: Update from Breast Cancer Surveillance Consortium, Radiology 283 (1) (2017) 59–69.
[39] R. Ben-Ari, A. Akselrod-Ballin, L. Karlinsky, S.Y. Hashoul, Domain specific convolutional neural nets for detection of architectural distortion in mammograms, in: IEEE Int. Symposium on Biomedical Imaging (ISBI), 2017, pp. 552–556.
[40] N. Wu, J. Phang, J. Park,..., Deep neural networks improve radiologists' performance in breast cancer screening, in: arXiv:1710.03778, 2019.



Ran Bakalo received the B.Sc. degree in computer science and the M.Sc. degree in computer science from University of Haifa. He is currently working as a researcher in the Medical Imaging Analytics group at IBM Research, Israel. His research interests include deep learning, machine learning, computer vision and medical imaging processing. His research is currently focused on classification and detection in digital breast tomosynthesis and mammograms.

Jacob Goldberger received the Ph.D. degree in 1998 from Tel-Aviv University, Israel, in electrical engineering. He was a post-doctoral fellow in the computer vision group at the Weizmann institute and later he was a post-doctoral fellow in the machine learning group at the University of Toronto. In 2004 he joined the engineering faculty at Bar-Ilan University where he is now a Professor. His research deals with developing and analyzing efficient statistical algorithms for learning and inference in the context of classical machine learning tasks such as classification, clustering and embedding and applying these algorithms to a large variety of applications such as computer vision, speech processing, medical imaging and natural language processing. In recent years his research is focused on addressing these challenges in the context of deep learning.

Rami Ben-Ari is a research staff member and technical lead for computer vision and deep learning at Video-AI technologies in IBM- Research, Haifa Lab, and is an adjunct professor at Bar-Ilan university, faculty electrical and computer engineering. He holds a BSc and MSc in Aerospace engineering from Technion, Israel Institute of Technology and a PhD in Applied Mathematics from Tel-Aviv University in computer vision. His research interests cover medical image analysis and more recently video understanding, including action recognition and self-supervised learning.