Multi-scale Graph Neural Networks for Mammography Classification and Abnormality Detection

Guillaume Pelluet, Mira Rizkallah, Mickael Tardy, Oscar Acosta, Diana Mateus

Multi-scale Graph Neural Networks for Mammography Classification and Abnormality Detection

Guillaume Pelluet[1,2;3][0000−0002−6537−6768], Mira Rizkallah[1][0000−0001−7724−9304], Mickael Tardy[2][0000−0003−4069−9517], Oscar Acosta[3][0000−0002−5447−1479], and Diana Mateus[1][0000−0002−2252−8717]

[1] Ecole Centrale de Nantes, LS2N, UMR CNRS 6004, Nantes, France
[2] Hera-MI, SAS, Nantes, France
[3] Université de Rennes 1, LTSI, UMR 1099, Rennes, France

Abstract. Early breast cancer diagnosis and lesion detection have been made possible through medical imaging modalities such as mammography. However, the interpretation of mammograms by a radiologist is still challenging. In this paper, we tackle the problems of whole mammogram classification and local abnormality detection, respectively, with supervised and weakly-supervised approaches. To address the multi-scale nature of the problem, we first extract superpixels at different scales. We then introduce graph connexions between superpixels (within and across scales) to better model the lesion's size and shape variability. On top of the multi-scale graph, we design a Graph Neural Network (GNN) trained in a supervised manner to predict a binary class for each input image. The GNN summarizes the information from different regions, learning features that depend not only on local textures but also on the superpixels' geometrical distribution and topological relations. Finally, we design the last layer of the GNN to be a global pooling operation to allow for a weakly-supervised training of the abnormality detection task, following the principles of Multiple Instance Learning (MIL). The predictions of the last-but-one GNN layer result in a superpixelized heatmap of the abnormality probabilities, leading to a weakly-supervised abnormality detector with low annotations requirements (i.e., trained with image-wise labels only). Experiments on one private and one publicly available datasets show that our superpixel-based multi-scale GNN improves the classification results over prior weakly supervised approaches.

Keywords: Mammography · Superpixels · Graph · GNN · Classification · Detection · Segmentation

1 Introduction

Breast cancer is the most common cancer in women worldwide [7]. Early-stage screening through mammography has demonstrated strong efficacy in reducing mortality caused by breast cancer. The detection of abnormal areas is a key step

in the diagnosis process. However, since mammograms are 2D X-ray projections, the abnormal lesions can be overshadowed by superimposing high-density tissues.

Current deep learning methods built on Convolutional Neural Networks (CNN) have demonstrated good performances in the automated analysis of individual mammograms, e.g., for the tasks of malignant image or region classification [1, 3]. To improve the ability to detect small lesions (e.g., calcifications) at higher resolutions, a common alternative are patch-wise classification or detection approaches [16, 15]. For instance, Shen et al. [16] proposed converting a patch classifier into a whole image classifier by modifying the last layers of the network. Instead, Ribli *et al.* [15] opt for an object detector approach (Faster RCNN [8]). Fully supervised patch-wise approaches such as [16, 15] require region-wise delineations of the lesions, which are not part of clinical protocols. Removing the need for lesion delineations, Choukroun *et al.* [4] proposed a weakly-supervised Multiple Instance Learning (MIL) approach, where the model for patch predictions is trained from image-wise labels only. Our approach is also weakly supervised but we rely on superpixels instead of patches which later allows for detailed abnormality region segmentation. To cope with the variability of both lesion's size and shape, we rely on a multi-scale graph representation of the mammogram where each node represents a superpixel at a specific scale, and each superpixel regroups neighboring pixels sharing common characteristics (e.g., pixel intensity) [2].

Despite mammograms being 2D uniform grids, abnormalities do not appear at a single scale, nor are uniformly distributed in the Euclidean space. Motivated by those two facts, we introduce an alternative representation of the image based on a multi-scale graph and model the classification task with a Graph Neural Network (GNN).

Graphs are powerful representation tools used to model the structure of the underlying domain in medical imaging [10]. GNNs contextualize patch-level information from their neighbors through message passing, thus enabling learning from both individual node features and the graph topological information. Few recent works have addressed the analysis of mammographic images with GNNs. In Du *et al.* [6], the authors introduced the fully supervised Graph Attention Networks (GAT) for mono-view mammography image classification. Their model relies on a multi-scale graph representation of the mammogram, where each node corresponds to a squared patch in a specific scale and zooming operations from radiologists are modeled as connections between neighboring scales. Graphs have also been useful for modeling intrinsic geometric and semantic relations between ipsilateral views (Liu *et al.* [12]).

In this paper, we propose the supervised learning of the mono-view mammogram classification task and the weakly-supervised learning of the abnormality detection task, both based on a single multi-scale graph and a Graph Convolutional neural Network (GCN) which only needs image-wise ground truth class labels for training. Unlike Du *et al.* 's graph [6], where relationships between scales are independent, our graph draws connections within and across scales,

i.e., between each superpixel and its spatial neighbors in the same scale and between neighboring superpixels in different scales.

In practice, our method assigns a set of features for each node in the multi-scale graph by applying a customized encoder (denoted as Backbone). The multi-scale graph and the node features are then fed to a GCN which outputs a global classification of the mammogram along with multi-scale heat-maps used for lesion detection by adaptive thresholding.

The model is trained on an in-house private dataset (**PRV**) and then evaluated on both the private dataset and a public dataset (**INB**), both of them consisting of mammograms from different populations, countries of origin, and acquired with different mammography system vendors. The experimental validation shows that our proposed method yields competitive global classification results while outperforming state-of-the-art weakly-supervised methods for lesion detection on an unseen manufacturer dataset. To the best of our knowledge, our learning framework scheme is the first weakly supervised method reaching an AUC score of around 0.83 for breast-wise classification.

2 Methods

Let a breast imaging exam be composed of a mammogram $\mathcal{I}$, corresponding to a Craniocaudal (CC) or a Medio-lateral oblique (MLO) view. Our goal is to perform a breast cancer screening classification, intended to capture the presence of malignant regions on the mammogram. We treat a mammogram as benign when it has no or benign lesions only. We consider an image as malignant if it contains at least one malignant lesion. In this context, we propose a deep learning framework, as depicted in Figure 1. The framework takes as input a breast image (mammogram) and an approximate range of possible lesion's sizes, then outputs a prediction of the probability of the presence of a malignant region in the image, and a region-wise prediction on different scales useful for lesion detection but also improving the interpretability of the model.

The framework is composed of three independent modules: the first module consists of a multi-scale over-segmentation of the mammogram based on the superpixels and a multi-scale graph generation. The second block is the feature extraction module, which computes the feature vectors of the nodes, i.e., the features assigned to each superpixel. Finally, the last module is a GCN taking as input both the node features and the multi-scale graph to output the probability of malignancy for every node and the whole mammogram. In the following, we give a detailed description of the three modules.

2.1 Multi-scale graph generation

Multi-scale oversegmentation To allow for better detection of malignant regions with variable sizes, we adapt the method proposed in Hang. *et al.* [9] to over-segment the mammogram $\mathcal{I}$ at several scales into superpixels. To generate the region candidates for each scale, we rely on a modified version [14]

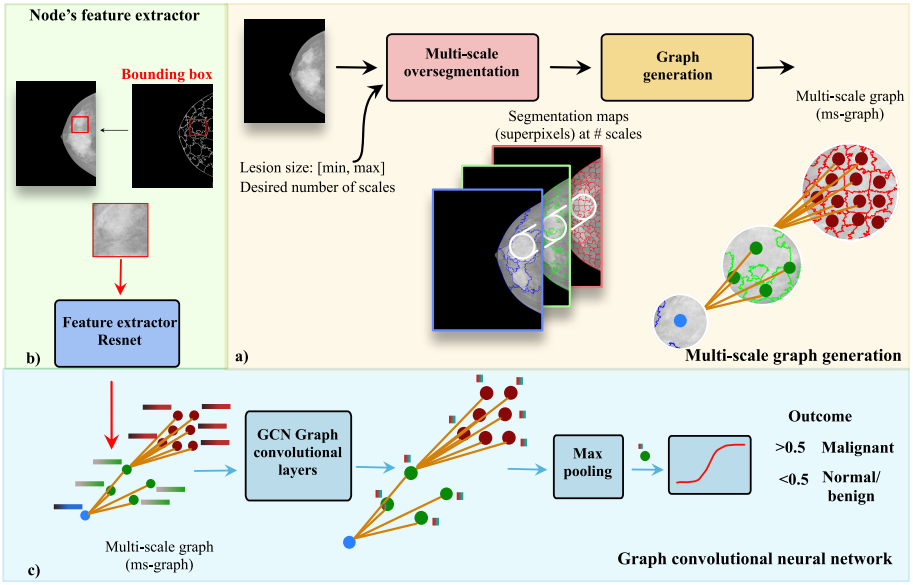


Fig. 1: Overview of the proposed learning framework consisting of three modules: a multi-scale graph generation module (a), a feature extraction module (b) and a graph convolutional neural network module (c). The framework requires as input a mammogram, the lesion's size range and the number of desired scales as input, and outputs a probability of malignancy for the whole image and an associated region-wise heatmap.

of the Scalable Simple Linear Iterative Clustering (SSLIC) algorithm proposed by Lowekamp *et al.* [13]. This modified version allows us to modulate the compactness parameter according to the variance of the superpixel features. As a result of this step, we obtain a multi-scale clustering $\mathcal{S} = \{S_1, S_2, \ldots, S_M\}$, with S_M the superpixel oversegmentation at scale m and M the number of scales. Equivalently, $\mathcal{S}$ can be seen as a collection of N superpixels s_i (with different sizes and shapes) resulting from all the scales such that $\mathcal{S} = \{s_i\}_{i=1}^{N}$.

Graph generation In order to capture intra-scale and inter-scale relationships in the mammogram $\mathcal{I}$, i.e., between sub-regions inside a specific scale and between corresponding superpixels in different scales respectively, we build a multi-scale graph. More precisely, given the image $\mathcal{I}$ and its multi-scale superpixel clustering $\mathcal{S}$, we build a graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ consisting of a finite set $\mathcal{V}$ of vertices and a set of edges $\mathcal{E}$. Each vertex in $\mathcal{V}$ corresponds to a superpixel $s_i \in \mathcal{S}$, therefore $\|\mathcal{V}\| = N$. To build $\mathcal{E}$, we connect each vertex in $\mathcal{V}$ with its 4 neighbors in the spatial (2D) domain (i.e., top, bottom, left, and right), and

its 2 neighbors across scales (i.e., nearest smaller and the nearest larger scales). An illustration of the resulting multi-scale (ms-graph) is shown in Figure 1.a. A binary adjacency matrix $\mathbf{A} \in [0,1]^{N \times N}$ is associated to the multi-scale graph, such that each entry $a_{i,j}$ is set to 1 if there is an edge connecting the two vertices s_i and s_j, otherwise, $a_{i,j}$ is set to 0.

2.2 Node Features Extraction

Furthermore, we extract relevant features for each node in the graph i.e., for each superpixel s_i, and store them in the i^{th} row $\mathbf{x}_i$ of a feature matrix $\mathbf{X}$. We use a weakly supervised Resnet22 [19], trained on mammograms with a MIL approach, to extract features vectors $\{\mathbf{x}_i\}$. In the rest of the paper, we denote this network as the Backbone $\mathcal{B}$. To apply $\mathcal{B}$ to a node, we first compute a bounding-box around each superpixel s_i. We then resize the extracted patch to fit the input size of the backbone network (i.e., the size of the original training patches). At the output of this module, we get the node features matrix $\mathbf{X} \in \mathbb{R}^{N \times D_{in}}$, where D_{in} is the dimensionality of the feature vector of each node.

2.3 Graph Convolutional Neural Network

Once the multi-scale graph is built, and the extracted features assigned to the nodes, we aim at exploiting the relationships between superpixels along with their deep features to classify the mammogram and provide an output class (benign or malignant). To do so, we rely on a GCN fed with the graph $\mathcal{G}$ as shown in Figure 1.

From the architecture standpoint, the network is composed of four graph convolutional layers. Each layer GCN_n (with $n \in \{1, \ldots, 4\}$) is composed of a graph convolutional operator, as defined in Kipf *et al.* [11], followed by an activation function and a dropout layer. Each GCN_n gets as input the feature matrix from the previous layer $\mathbf{H}_{n-1}$ ($\mathbf{X}$ for the first layer) and provides as output the transformed matrix $\mathbf{H}_n$. The final output node feature matrix $\mathbf{H}_{out} = \mathbf{H}_4 \in \mathbb{R}^{N \times D_{out}}$ contains the node representation encoding both graph structural properties and node features. In our case, we fix D_{out} to 1 for binary malignancy probability. The feature matrix $\mathbf{H}_{out}$ is aggregated with a global maximum graph pooling layer, retaining only the node with the maximum value h_{max}. A non linear activation layer (a sigmoid function denoted as $\sigma(\cdot)$) is then applied on h_{max} to obtain the probability of malignancy. The prediction for the entire image is computed as:

$$\hat{\mathbf{y}} = \sigma(f_\theta(h_{max})) \tag{1}$$

where f_θ represents the whole GCN architecture with parameters θ trained with an image-wise weighted cross-entropy loss.

2.4 Implementation details

The weakly-supervised backbone encoder is trained using a Resnet22 architecture respecting the MIL training strategy. In order to compute the features of a

specific node in the graph, we use the backbone encoder and extract 256 features from the second but last layer. Our Multi-Scale Graph Convolutional Network ($\mathcal{MSGCN}$) model is trained following the same strategy (i.e., MIL) using the library DGL [20] with PyTorch [5] in backend. The training was performed using Adam optimizer with an initial learning rate of $5 \cdot 10^{-4}$ and default parameters. For the graph convolutional layers, we used a weight decay factor of $1 \cdot 10^{-5}$. All the experiments were trained for at least 10000 epochs. While training the GCN, we apply a dropout with a probability of 0.1 at each layer and we used a batch size of 64. The model was trained on NVIDIA A100 GPU.

The input to our framework is a pre-processed mammogram. The preprocessing consists of the following steps: the right breasts are horizontally flipped to align the breast to the left of the image. The background is cleaned to remove labels using triangle thresholding; the cleaned mammograms are then cropped to the bounding box around the breast to avoid using background information; the cropped images are resized to a height of 3072 pixels. To increase the contrast of the resized mammogram, we perform histogram stretching between the intensities corresponding to the 2^{nd} percentile and the 99^{th} percentile. We finally normalize the intensity values between 0 and 1.

3 Experimental results

3.1 Experimental setup

Datasets Experiments are performed on mammograms originating from different populations, locations (countries), and mammography system vendors. More precisely, we evaluate our model on two different datasets: a private dataset managed in-house, and a public dataset. The former, denoted as **PRV**, is composed of 3162 Full Field Digital Mammography (FFDM) images from four different vendors, namely Fujifilm, GE, Hologic, and Planmed. For all the malignant mammograms, pixel-level annotations of the lesions, drawn by the clinical experts, are provided (only used for evaluation). The publicly available INbreast dataset is composed of 410 FFDM images from the Siemens mammography system. Similar to **PRV**, images have pixel-level annotations for each lesion delineated by an expert radiologist. In the following, we refer to this dataset as **INB**. The distributions of the two classes (benign/normal or malignant) in both datasets are given in Table 1.

Dataset	Samples	Benign	Malignant	Train set	Test set
PRV	3162	1597 (50.5 %)	1565 (49.5 %)	2658	504
INB	410	310 (75.6 %)	100 (24.4 %)	0	410

Table 1: Composition of the datasets

In order to evaluate the proposed approaches, we used **PRV** for training and testing while keeping the same samples in the test set as done in the baseline. We split the remaining train set into 80-20 train/validation splits while keeping the validation set balanced. We used the full **INB** dataset for evaluation as well as a subset defined in [18] for a fair comparison to similar works.

The superpixels are generated at 4 different scales for both datasets, and their statistics are given in Table 2, and shown in Figure 2.

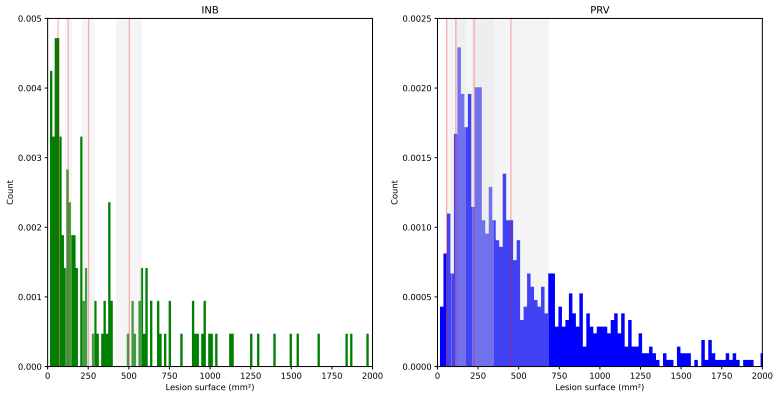


Fig. 2: lesion's size's distribution in both datasets: **INB** and **PRV**. The red vertical lines and gray zones correspond to the average and standard deviation (in mm^2 of superpixel's size in each scale respectively)

	PRV	**INB**
Scales	Average size (std)	
0	56.57 (29.59)	63.36 (9.73)
1	113.09 (59.09)	126.52 (19.39)
2	226.15 (117.75)	252.04 (38.59)
3	452.17 (234.87)	502.67 (76.96)

Table 2: Average and standard deviation (in mm^2) of the superpixel's size in each scale in both datasets.

Evaluation protocol We evaluate the performance of our learning framework for three tasks: global classification (image-wise and breast-wise), lesion detection, and lesion segmentation. We used breast-wise, image-wise, and pixel-level

ground-truths to evaluate the proposed model for breast-wise / image-wise malignancy classification and local detection/segmentation. To evaluate our model at a breast-level, we average breasts' malignancy predictions when the breast is composed of several mammograms.

The Area Under the Curve (AUC) was used as a metric to assess the performance for image-wise classification. As for the local detection assessment, the Area under the Free-Response ROC Curve (AUFROC) and the TPR@FPPI metric were used. Dice Score was used to measure the malignant lesions segmentation performance.

3.2 Evaluation of the proposed framework

In this section, we focus on the analysis of the performance of our learning scheme in the context of two tasks: the global image/breast classification and abnormal region detection tasks.

Ablation study of multi-scale features In order to show the interest of the multi-scale representation, we start with an ablation study of the scales parameter given as input to our learning framework. More precisely, choosing one or more scales implies generating a one-scale graph or a multi-scale graph respectively. Node features are extracted accordingly. In each case, both the graph and the node features are fed to the GCN.

In Table 3, we report the image-wise classification AUC obtained with varying scale parameters. We observe that the scale 0 and 3 have the lowest AUC for image-wise classification, while the performance using scales 1 or 2 reaches 0.77. The difference in terms of performance can be explained by the fact that superpixels in scales 0 and 3, with an average isotropic size of around 175 and 479 pixels (Table 2) respectively, had to be resized before being fed to the feature extractor module. The sub-optimal resizing procedure generates blur or noise artifacts in the interpolated patch. Moreover, exploiting the features originating from scale 1 and scale 2 simultaneously (i.e., 1&2) improves the classification performance to reach an AUC of 0.80, similarly to the performance with scales 1&2&3. This shows that the GCN is able to mix the information originating from multiple scales, leading to an improvement in the classification performance. With no benefit observed when adding scale 3, for the rest of the analysis, we will focus on scales 1 and 2 to generate the graph and node features.

Scales	0	1	2	3	1&2	1&3	2&3	1&2&3
Image-wise AUC	0.57	0.77	0.77	0.65	**0.80**	0.65	0.73	**0.80**

Table 3: Scales ablation with $\mathcal{MSGCN}$ on **PRV** dataset.

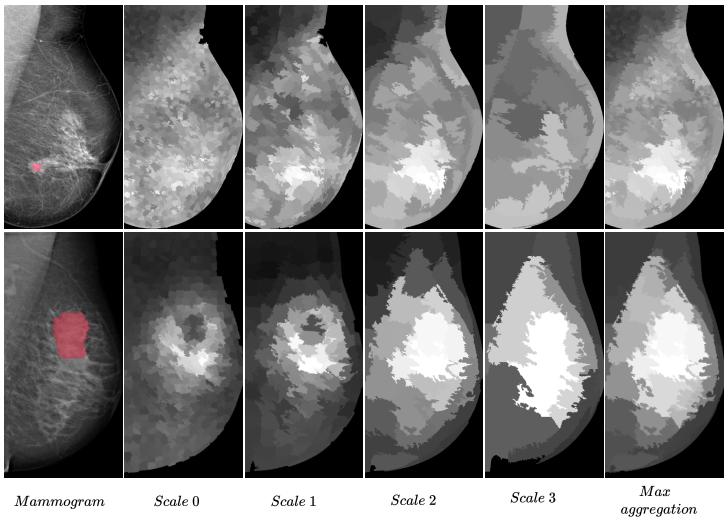


Fig. 3: Illustration of two types of lesions with different sizes (a calcification cluster in the first row, and a mass in the second row), along with their respective heatmaps generated at different scales and the aggregated $\mathcal{MSGCN}$ heat-map.

Patches vs Superpixels vs MSGCN Having fixed the scales to consider in $\mathcal{MSGCN}$, we show here the interest of each of the modules in our framework by comparing the classification and detection performances against 2 related schemes:

- The Backbone ($\mathcal{B}$) scheme, based on the Backbone network, fed with square patches of size 256×256 and a stride of 128. This approach uses neither superpixels, the multi-scale graph, nor the GCN.
- The Superpixel-based Backbone ($\mathcal{BSP}$) scheme where we rely on the features extracted from the superpixels at two scales (1&2) (with our multi-scale graph generation module) and perform direct whole image classification, without entering the GCN module.

For the evaluation of the detection task, we rely on the intermediate computation of activation maps. Our $\mathcal{MSGCN}$ leverages the topological and textural information of the multiple scales to provide a node embedding that is suitable for both classification and detection tasks. Indeed, by design, the model does not only provide a malignancy probability for the whole image but also hands over a probability for each node in the multi-scale graph, yielding a set of activation maps $\mathcal{M}_G^{(m)}$ with m being the scale. We create the final activation map $\mathcal{M}_G$ for an image, by aggregating the activation maps with a maximum pooling across scales. To each pixel on the mammogram, we assign the maximum activation obtained for this pixel over the whole set $\mathcal{M}_G^{(m)}$. Examples of the resultant activation maps for two lesions are shown in Figure 3. The superpixel-based scheme

$\mathcal{BSP}$ provides also a set of activation maps $\mathcal{M}_S^{(m)}$ that are also aggregated into a single map $\mathcal{M}_\mathcal{S}$ with a similar max pooling operation. The patch-based scheme $\mathcal{B}$ also gives, by design, the direct activation map for an image, corresponding to the patch-wise predictions before the last aggregation step of $\mathcal{B}$. Activation maps for $\mathcal{M}_\mathcal{S}$ and $\mathcal{M}_B$ can be seen in Figure 4.

From the above activation maps, we can detect the malignant regions using an adaptive thresholding procedure. In Table 4, we can see the results of image-wise, breast-wise classification, and abnormality detection obtained when using the three schemes: $\mathcal{B}$, $\mathcal{BSP}$ and our Multi-scale Superpixels Graph Neural Network denoted $\mathcal{MSGCN}$.

Classification Table 4 shows that $\mathcal{MSGCN}$ allows a better classification (image-wise and breast-wise) compared to the backbone approach $\mathcal{B}$ on the public dataset **INB**. Indeed, in Pelluet *et al.* [14], we had shown that superpixels are suitable for a finer segmentation of lesions and thus we expect features extracted from the finer segmentation better capture the information about the lesion. Moreover, GNNs are efficient in summarizing and propagating information between different scales. The simpler patch-based backbone patches approach $\mathcal{B}$ gives a high AUC for global classification at the expense of lower detection performance.

Detection In order to evaluate the detection performance i.e., the ability to detect and accurately localize malignant lesions, we evaluated the predicted heatmaps using the FROC curve on all findings (excluding distortions) on **INB**. To plot the curves, we applied thresholds on the probability values of the heatmaps $\mathcal{M}_B$, $\mathcal{M}_S$ and $\mathcal{M}_G$. Figure 4 shows the activation maps obtained with the 3 approaches : $\mathcal{M}_B$, $\mathcal{M}_S$, and $\mathcal{M}_G$. In the case of the method $\mathcal{B}$, only one activation map is provided. This is not optimal knowing the statistical distribution of the lesion's size shown in Figure 2. Instead, aggregating the $\mathcal{MSGCN}$ heatmaps across scales provides a better detection, having learned the embedding of superpixels features at different scales simultaneously. Table 4 shows an improvement of the TPR, from 0.52 when using $\mathcal{M}_B$, to 0.73 with $\mathcal{M}_S$, at the expense of a higher FPPI. The best detection performance is obtained with $\mathcal{MSGCN}$ ($\mathcal{M}_G$) which reaches a TPR@FPPI of 0.98@1.01 increasing the initial backbone's TPR and FPPI by 88.4% and 71.2%, respectively. The increase of the FPPI can be explained by the communication of bad predictions between neighboring nodes in the multi-scale graph.

Comparative performance analysis against state-of-the-art methods We evaluate our methods against three state-of-the-art learning methods, two of which are fully supervised approaches [15, 17] and one is a weakly supervised method [21]. The results are shown in Table 5.

- For the *abnormality detection task*, evaluation is only performed on the malignant images of the **INB** dataset for a fair comparison with the state-of-the-art. To generate the results for Shen *et al.* [17], the publicly available

Method	Test set	Supervision	Train data	TPR@FPPI	Image-Wise AUC	Breast-Wise AUC
$\mathcal{B}$	F	W	Private	0.52@0.59	0.8	0.8
$\mathcal{BSP}$	F	W	Private	0.73@0.74	0.80	0.82
$\mathcal{MSGCN}$	F	W	Private	**0.98@1.01**	**0.82**	**0.83**

Table 4: Performance of $\mathcal{MSGCN}$, $\mathcal{B}$ and $\mathcal{BSP}$ in terms of image-wise and breast-wise classification on the full **INB** dataset.

model was used. The TPR@FPPI was recomputed using their top 2% pooling, as suggested in the original paper.

– For the *classification task*, results of [21] are taken from [18] where they were also evaluated on the same subset of the **INB** dataset. The full dataset **INB** is used for evaluation against the fully supervised learning method [15]. For a fair comparison with the work of Wu et al. [21], Shen et al. [17], and Ribli et al. [15], we compute breast-wise AUC on an image subset of the test dataset **INB**.

Method	Test set	Supervision	Train data	TPR@FPPI	Breast-Wise AUC
Wu [21]	S	Fully	Private	NA	0.80
Ribli [15]	S	Fully	Private & DDSM	NA	**0.97**
$\mathcal{MSGCN}$ (ours)	S	W	Private	NA	0.90
Ribli [15]	F	Fully	Private & DDSM	0.90@0.30	**0.95**
Shen [17]	F	W	Private	0.97@1.94	0.82
$\mathcal{MSGCN}$ (ours)	F	W	Private	**0.98@1.01**	**0.83**

Table 5: The performance of our scheme compared to state-of-the-art methods on the INBreast dataset. For the test set: F corresponds to the full INBreast dataset, and S is for the test subset defined in Stadnick *et al.* [18]. As for the supervision, the methods are fully supervised (Fully) if they need pixel-wise ground truth labels for training. They are weakly supervised (W) if they only require only image-wise ground truth labels.

Our method shows better results when compared to Wu et al.[21], with an improvement of 11.25% on the breast-level AUC while our method is weakly supervised. The best performance on the **INB** subset is yielded by the fully-supervised method of Ribli et al.[15], with a breast-wise AUC of 0.97.

Furthermore, the proposed method $\mathcal{MSGCN}$ improves the detection performance with a TPR@FPPI of 0.98@1.01, in comparison to other weakly and fully supervised methods, as shown in Table. 5 and Figure 5. However, Ribli *et al.* [15] achieves the lowest FPPI=0.3, yielding a very low number of false positives. Finally, while the fully supervised model of [15] outperforms all the other methods with a breast-wise AUC of 0.95, our method performs relatively well with a

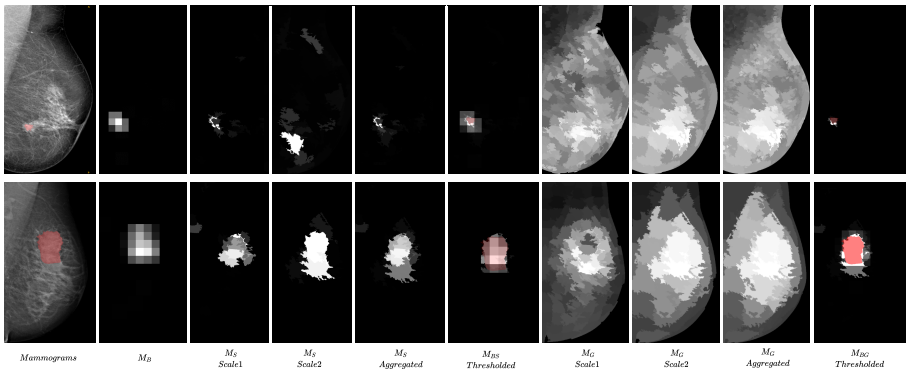


Fig. 4: Two examples of segmentation maps obtained from the merged heatmaps $\mathcal{M}_{BS}$ and $\mathcal{M}_{BG}$ with a threshold of 0.5. The second column corresponds to the Backbone heatmap extracted from the Backbone network. The 6^{th} column corresponds to the merged heatmap resulting from aggregating $\mathcal{M}_B$ and $\mathcal{M}_S$ *Aggregated* (i.e., the 2^{nd} and the 5^{th} column). The last column illustrates the resulting heatmap after aggregating $\mathcal{M}_B$ and $\mathcal{M}_G$ (i.e., the second and the 9^{th} columns).

TPR@FPPI of 0.98@1.01 on the **INB** among the three weakly supervised methods (relying only on whole-image labels), with a higher TPR and fewer false negatives. This shows that our model is generalizable to datasets from other manufacturers.

3.3 A Finer Lesion Segmentation with Superpixels

To further illustrate the interest of using a superpixel-based segmentation, we evaluated the lesion segmentation which we can obtain by applying a merging operation on activation maps as explained below. More precisely, to exploit the superpixel ability to adhere to object boundaries for lesion segmentation, we generated two new activation maps $\mathcal{M}_{BG}$ and $\mathcal{M}_{BS}$, by merging the information from $\mathcal{M}_B$ with $\mathcal{M}_G$ or $\mathcal{M}_S$ (e.g. Figure 4, 5^th and 9^th column) respectively. This is performed using a simple average aggregation as shown in Eq. 2.

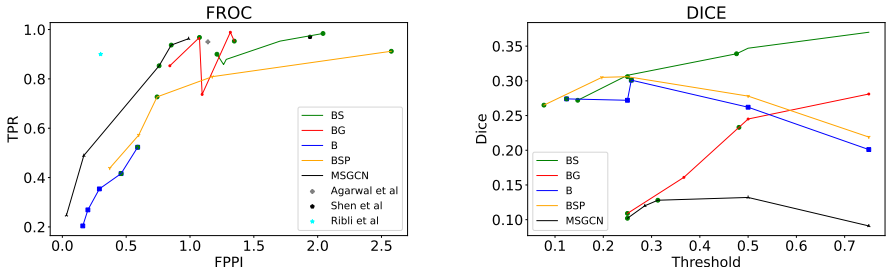
$$\begin{aligned} \mathcal{M}_{BG} &= \frac{sum(\mathcal{M}_B, \mathcal{M}_G)}{2} \\ \mathcal{M}_{BS} &= \frac{sum(\mathcal{M}_B, \mathcal{M}_S)}{2} \end{aligned} \tag{2}$$

In Figure 4, we provide two examples of the obtained heatmaps for two different lesion's sizes. We can see that with heatmaps ($\mathcal{M}_{BG}$ and $\mathcal{M}_{BS}$), the merged information retains the boundaries of the superpixels which leads to a better loyalty to the lesion borders. This performance is also noticeable looking at the DICE score in Table 5. We can see that a better performance in terms

of segmentation for all malignant lesions is obtained using our merged heatmap $\mathcal{M}_{BS}$ with Dice Scores of 0.37, outperforming the backbone $\mathcal{B}$ and the backbone applied to superpixels bounding-boxes $\mathcal{BSP}$ which both have a DICE of 0.30. Although the $\mathcal{MSGCN}$ brings a better detection performance, it has a lower DICE score of 0.13. Indeed, the aggregation of the different scales appears to be not optimal for segmentation. This issue will be tackled in a future investigation.

Methods	Supervision	Train data	TPR@FPPI	Dice
Shen [17]	W	Private	0.97@1.94	0.34
Agarwal [3]	Fully	OPTIMAM	0.95@1.14	NA
$\mathcal{B}$ (ours)	W	Private	0.52@0.59	0.30
$\mathcal{BSP}$ (ours)	W	Private	0.73@0.74	0.30
$\mathcal{MSGCN}$ (ours)	W	Private	**0.98@1.01**	0.13
$\mathcal{BG}$ (ours)	W	Private	0.94@1.03	0.28
$\mathcal{BS}$ (ours)	W	Private	0.93@1.03	**0.37**

Table 6: Segmentation methods on **INB**



(a) FROC (b) DICE Score

Fig. 5: FROC curve and Dice Score representation of detection performance on the full INBreast dataset.

4 Conclusion

In this work we proposed a new learning framework for mammography classification and lesion detection/localization based on graph neural networks. We build our method on top of a backbone feature extraction module, and then improve its reliability in terms of classification, detection, and segmentation. To do so, we rely on three modules i) a novel multi-scale graph representation of the mammogram to model the zoom-in radiologist operation, allowing the backbone model to better capture relevant information at different scales; ii) deep features obtained for superpixels at different scales are then fed along with the graph; and iii) graph neural network to enable message passing between superpixels within

and between scales. The lesion segmentation performance is improved when using superpixels which adhere well to object boundaries. Our weakly-supervised method based on a multi-scale graph improves the classification and detection results over the patch-based baseline, and compares well to state-of-the-art approaches, which do not consider graphs.

It is worth noting that the actual proposed model ($\mathcal{MSGCN}$) propagates the information uniformly through the local neighborhood without taking into account individual pair-wise correlations which can be different between neighboring superpixels. In order to improve the performance of the model, we will consider adding weights to the graph edges in future work.

References

1. Abdelrahman, L., Al Ghamdi, M., Collado-Mesa, F., Abdel-Mottaleb, M.: Convolutional neural networks for breast cancer detection in mammography: A survey. Computers in Biology and Medicine **131**, 104248 (2021). https://doi.org/https://doi.org/10.1016/j.compbiomed.2021.104248, https://www.sciencedirect.com/science/article/pii/S0010482521000421
2. Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Süsstrunk, S.: Slic superpixels. Technical report, EPFL (06 2010)
3. Agarwal, R., Díaz, O., Yap, M.H., Lladó, X., Martí, R.: Deep learning for mass detection in Full Field Digital Mammograms. Computers in Biology and Medicine **121**, 103774 (jun 2020). https://doi.org/10.1016/j.compbiomed.2020.103774
4. Choukroun, Y., Bakalo, R., Ben-ari, R., Askelrod-ballin, A., Barkan, E., Kisilev, P.: Mammogram Classification and Abnormality Detection from Nonlocal Labels using Deep Multiple Instance Neural Network. Tech. rep. (2017). https://doi.org/10.2312/VCBM.20171232
5. Collobert, R., Kavukcuoglu, K., Farabet, C.: Torch7: A matlab-like environment for machine learning. In: BigLearn, NIPS Workshop (2011)
6. Du, H., Feng, J., Feng, M.: Zoom in to where it matters: a hierarchical graph based model for mammogram analysis (Cc) (2019), http://arxiv.org/abs/1912.07517
7. Ferlay J, Ervik M, L.F.C.M.M.L.P.M.Z.A.S.I.B.F..: Global cancer observatory: Cancer today. lyon, france: International agency for research on cancer https://gco.iarc.fr/today
8. Girshick, R.: Fast r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (December 2015)
9. Hang, M., Chandra, S., Crozier, S., Bradley, A.: Multi-scale sifting for mammographic mass detection and segmentation. Biomedical Physics & Engineering Express **5** (01 2019). https://doi.org/10.1088/2057-1976/aafc07
10. He, Y., Zhao, H., Wong, S.T.: Deep learning powers cancer diagnosis in digital pathology. Computerized Medical Imaging and Graphics **88**, 101820 (2021). https://doi.org/https://doi.org/10.1016/j.compmedimag.2020.101820, https://www.sciencedirect.com/science/article/pii/S0895611120301154
11. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. CoRR **abs/1609.02907** (2016), http://arxiv.org/abs/1609.02907
12. Liu, Y., Zhang, F., Chen, C., Wang, S., Wang, Y., Yu, Y.: Act Like a Radiologist: Towards Reliable Multi-view Correspondence Reasoning for Mammogram Mass Detection. IEEE Transactions on Pattern Analysis and Machine Intelligence **8828**(c), 1–15 (2021). https://doi.org/10.1109/TPAMI.2021.3085783

13. Lowekamp, B.C., Chen, D.T., Yaniv, Z., Yoo, T.S.: Scalable simple linear iterative clustering (sslic) using a generic and parallel approach. Kitware, Inc. (2018)
14. Pelluet, G., Rizkallah, M., Acosta, O., Mateus, D.: Unsupervised multimodal supervoxel merging towards brain tumor segmentation (2021)
15. Ribli, D., Horváth, A., Unger, Z., Pollner, P., Csabai, I.: Detecting and classifying lesions in mammograms with Deep Learning. Scientific Reports **8**(1), 1–7 (2018). https://doi.org/10.1038/s41598-018-22437-z
16. Shen, L., Margolies, L., Rothstein, J., Fluder, E., McBride, R., Sieh, W.: Deep learning to improve breast cancer detection on screening mammography. Scientific Reports **9**, 1–12 (08 2019). https://doi.org/10.1038/s41598-019-48995-4
17. Shen, Y., Wu, N., Phang, J., Park, J., Liu, K., Tyagi, S., Heacock, L., Kim, S.G., Moy, L., Cho, K., Geras, K.J.: An interpretable classifier for high-resolution breast cancer screening images utilizing weakly supervised localization (feb 2020), http://arxiv.org/abs/2002.07613
18. Stadnick, B., Witowski, J., Rajiv, V., Chledowski, J., Shamout, F.E., Cho, K., Geras, K.J.: Meta-repository of screening mammography classifiers. CoRR **abs/2108.04800** (2021), https://arxiv.org/abs/2108.04800
19. Tardy, M., Mateus, D.: Leveraging Multi-Task Learning to Cope With Poor and Missing Labels of Mammograms. Frontiers in Radiology **1**, 19 (jan 2022). https://doi.org/10.3389/fradi.2021.796078
20. Wang, M., Zheng, D., Ye, Z., Gan, Q., Li, M., Song, X., Zhou, J., Ma, C., Yu, L., Gai, Y., Xiao, T., He, T., Karypis, G., Li, J., Zhang, Z.: Deep graph library: A graph-centric, highly-performant package for graph neural networks. arXiv preprint arXiv:1909.01315 (2019)
21. Wu, N., Phang, J., Park, J., Shen, Y., Huang, Z., Zorin, M., Jastrzebski, S., Fevry, T., Katsnelson, J., Kim, E., Wolfson, S., Parikh, U., Gaddam, S., Lin, L.L.Y., Ho, K., Weinstein, J.D., Reig, B., Gao, Y., Toth, H., Pysarenko, K., Lewin, A., Lee, J., Airola, K., Mema, E., Chung, S., Hwang, E., Samreen, N., Kim, S.G., Heacock, L., Moy, L., Cho, K., Geras, K.J.: Deep Neural Networks Improve Radiologists' Performance in Breast Cancer Screening. IEEE Transactions on Medical Imaging **39**(4), 1184–1194 (apr 2020). https://doi.org/10.1109/TMI.2019.2945514