# Annotation-free deep-learning framework for microcalcifications detection on mammograms

Paul Terrassin, Mickael Tardy, Nathan Lauzeral, Nicolas Normand

▶ **To cite this version:**

Paul Terrassin, Mickael Tardy, Nathan Lauzeral, Nicolas Normand. Annotation-free deep-learning framework for microcalcifications detection on mammograms. SPIE Medical Imaging: Computer-Aided Diagnosis, Feb 2024, San Diego (California), United States. pp.29, ⟨10.1117/12.3008304⟩. ⟨hal-04636407⟩

**HAL Id: hal-04636407**

**https://hal.science/hal-04636407v1**

Submitted on 9 Jul 2024

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

HAL Authorization

# Annotation-free deep-learning framework for microcalcifications detection on mammograms.

Paul Terrassin[a, b], Mickael Tardy[a, b], Nathan Lauzeral[a], and Nicolas Normand[b]

[a]Hera-MI, SAS, Saint-Herblain, France
[b]Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000 Nantes, France

## ABSTRACT

Breast cancer detection at an early stage significantly increases the chances of recovery for patients. Mammography (MG) is one of the most popular non-invasive and high-resolution imaging allowing radiologists to depict early signs of the disease. Microcalcifications (MCs) often occupy less than 1mm in size and can represent a high risk of suspicion depending on the spatial distribution, morphology, and their evolution over time. Their detection is challenging both the clinicians and computer-aided detection tools. In this work, we propose a novel annotation-free framework designed specifically for the MCs detection and trained in a self-supervised manner thanks to the generation of synthetic MCs. Inspired by the UNet3+ architecture, we reduced its number of parameters to make it applicable in practice and added multi-scale features to enrich fine-grained details with more global context information. Both multi-channel segmentation and multi-class classification tasks are implemented in a multi-scale output approach to catch MC of various sizes. We perform a comparison with several state-of-the-art methods, including different flavors of ResNet-22, ConvNeXt, and UNet3+. An analysis of classification and segmentation performances has been done, using the Gradient-weighted Class Activation Mapping method to make classifiers visually explainable. In this study, we used two public datasets, INBreast and Breast MicroCalcifications Dataset for validation and test purposes. We achieved an AUC score of 0.93 in the characterization of malignant MCs while having a semantic segmentation precision of 0.70. To the best of our knowledge, we are the first study claiming segmentation performances on the BMCD dataset.

**Keywords:** Breast cancer, Deep-learning, Microcalcifications, Computer-Aided Detection, Annotation-efficiency

## 1. INTRODUCTION

Breast cancer is the most diagnosed type of cancer, estimated to be one-third of new cancer cases, and the second leading cause of death amongst women.[1] Mammography (MG) is one of the most common medical imaging techniques used in breast screening programs. It allows the detection of breast tumors at an early stage of the disease while being non-invasive. MG is therefore a crucial step, as it considerably increases the chances of survival.[2]

Breast screening programs generate a large amount of data obtained through different medical imaging modalities. Hence, radiologists have to review a considerable amount of imaging while keeping their attention to small areas that could be malignant lesions. Furthermore, these images (i.e., mammograms) are high-dimensional in which the smallest abnormalities such as micro-calcifications (MCs) could be found within the areas below 1 $cm^2$ (e.g., $100 \times 100$ pixels in $4000 \times 3000$ images).[3] These MCs are deposits of calcium appearing as small white spots rounded or with irregularities and may be the first sign of pathology detectable with MG.[4]

To help radiologists cope with their workload, Computer-Aided Diagnosis (CAD) tools are being developed. The modern tools are mainly based on deep-learning approaches and aim to collaborate with radiologists to increase lesion detection rates while reducing the workload on negative exams.[5,6] Providing image-wise-only prediction (e.g., malignancy risk) may not be sufficient to offer optimal guidance for health practitioners. Indeed, the localization of findings through segmentation or detection tasks is crucial to make the predictions more comprehensible for radiologists. Localization tasks, though, require pixel-wise labeling or bounding boxes for fully-supervised training. Nonetheless, collecting those annotations is a laborious and time-consuming task. Moreover, most of the time only medical reports are available, and collecting ground truths from experts is not easily feasible.

In this context, the main contribution of this study is a framework based on a novel deep convolutional neural network (CNN) architecture and a training strategy using image-level annotations only. First, we designed a novel deep CNN inspired by the UNet3+ proposed by Huang et al.[7] and specifically adapted it to the MC detection problem. Indeed, our architecture profits from multi-scale features mixing global context and fine-grained details for the optimal decision-making, aiming to catch MCs of various sizes and generate multi-resolution semantic segmentations.[8,9] The multi-task approach brings balanced competition between classification and segmentation tasks, leading to better performances.[10,11] Our training strategy only requires the MG exam malignancy risk, accessible in the patient's medical report, as the framework generates its own synthetic MCs with its associated pixel-wise ground truths.

We evaluate the performances of the proposed framework and compare them to several state-of-the-art methods: ResNet-22,[12] ConvNext,[13] and UNet3+.[7] We use two public datasets: INBreast[14] and Breast MicroCalcifications Dataset (BMCD).[15] To the best of our knowledge, we are the first to compute and present semantic segmentation metrics on the second dataset.

## 2. RELATED WORKS

In this section, we summarize state-of-the-art methods based on deep-learning approaches for global breast cancer detection and the specific problem of MCs. One of the most common approaches consists of processing smaller breast regions of interest (ROI), called patches, aiming to preserve breast resolution. Due to their tiny sizes (e.g., $100 \times 100$ pixels in $4000 \times 3000$ images),[3] keeping the high semantic details of the lesion is crucial to efficiently identify their morphology and associate a risk of malignancy.[16–18] On the contrary, degrading pixel information may lead to the creation of noise or bright artifacts that can be confused with MCs.

Training deep CNN for a single MC classification task is the simplest and easiest way to deal with the lack of expert pixel-wise annotations. Shen et al. proposed an image-wise multi-classifier including MCs, trained on ROI patches and benefiting from the share of weights and locality properties from fully convolutional networks.[18] Sakaida et al. designed a $224 \times 224$ patch classifier to capture MC's presence and tested various ResNet architectures.[16] Finally, Quintana et al. have studied the impact of patch sizes and MG resolutions on the classification of breast lesions, including masses and calcifications. The authors conclude that a single size or resolution is not optimal for catching all lesions.[17] They combined patch classifiers of several sizes to generate the global image risk of malignancy prediction. All those methods focus on the image classification task. However, the classification alone may not give details about the localization of lesions, leading to less interpretable results.

Some methods proposed in the literature rely on a multi-task approach providing additional information thanks to segmentation or detection tasks. First, they aim to classify the whole image according to the risk of malignancy and label breast pixels associated with pathologies. Ouyang et al. introduced a multi-task CNN for the classification and segmentation of MCs with a self-adversarial approach and an end-to-end training manner.[19] However, this work trains and tests the proposed framework on a private dataset composed of experts' bounding boxes, which is a scarce and hard-to-find resource. Zhang et al. and Gerbasi et al. trained several CNNs in cascade for a classification first, and a localization task then.[20,21] This strategy creates a difficult workflow as the training, validation, and test steps depend on the first iteration. Such a design leads to extended processing times and prevents from taking advantage of concurrent training on multiple task.[10,11]

## 3. METHODS

In this section, we introduce the proposed framework, illustrated in Fig. 1. The framework is composed of three steps: a synthetic lesions generator $\mathcal{G}(\cdot)$, a novel CNN architecture $\mathcal{C}(\cdot)$, and the losses computation $\mathcal{L}(\cdot)$.

### 3.1 Network architecture

Our novel CNN architecture is based on the UNet3+ architecture proposed by Huang et al.[7] designed for small medical images ($320 \times 320$ pixels). However, our patch-wise approach uses bigger patches to provide as much context as possible while keeping the capacity to extract fine-grained details. Hence, to improve the computation efficiency, the number of filters $f$ has been reduced by a factor of 2. Our CNN is composed of an encoder $\mathcal{E}(\cdot)$, a bottleneck $\mathcal{B}(\cdot)$, and a decoder $\mathcal{D}(\cdot)$. The encoder's design is identical to the method used by Huang et al.[7] while

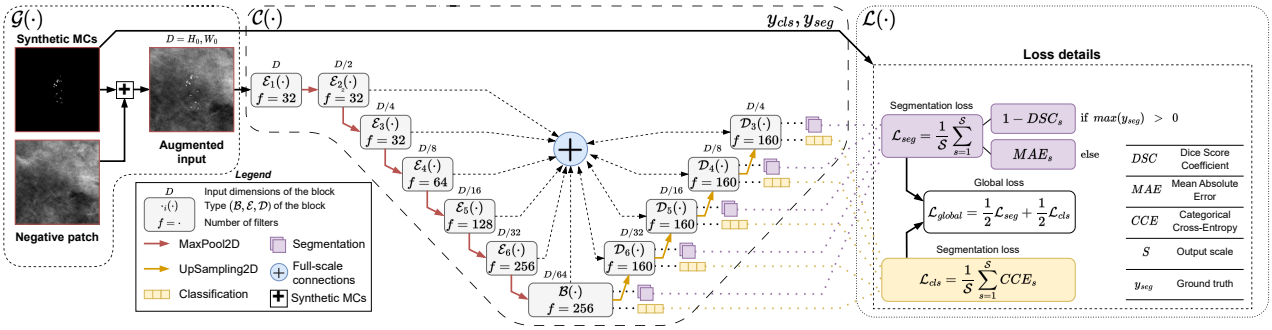


Figure 1. Overview of the proposed framework. Synthetic MCs are generated and inserted into a patch extracted from a negative exam and fed into the network. The CNN is based on multi-task and multi-scale approaches. Multi-classification and multi-channel segmentation tasks aim to distinguish soft tissues from benign and malignant MCs and are trained concurrently with $\mathcal{L}_{global}$ loss. $\mathcal{E}_n(\cdot)$, $\mathcal{B}(\cdot)$, and $\mathcal{D}_n(\cdot)$ denote the encoder, bottleneck and decoder at the $n_{th}$ scale with $f$ being number of filters.

the bottleneck uses the Scale Aware Pyramid Fusion (SAPF) module proposed by Feng et al.[22] SAPF aims to fuse features extracted through the encoder in a multi-scale approach. The decoder differs from the baseline, as we introduce Atrous Spatial Pyramid Pooling (ASPP) proposed by Chen et al.[23] which adds context for convolution operations with dilation aiming to enlarge receptive fields and catch multi-scale objects.

Moreover, the multi-task and multi-scale output approaches have been implemented to take advantage of both classification and segmentation tasks. First, the multi-class classification distinguishes soft tissues, benign calcified, and malignant MCs while the multi-channel segmentation task generates separate benign and malignant MC pixel-level prediction. Both tasks are designed to detect calcified areas and characterize them according to their risk of malignancy. To help the detection of MC of various sizes, a multi-scale output approach is used by generating multi-task predictions from the bottleneck and from each level of decoder.

All novelties introduced in this framework aim to generalize the multi-scale approach introduced by Huang et al. with full-scale skip connections.[7] The proposed CNN efficiently combines multi-scale context, using the finer-grained details at the top level and deeper features from the bottleneck. We designed our solution for the specific MC detection problem, segmenting and characterizing lesions thanks to the classification task that distinguishes benign from malignant calcifications.

### 3.2 Framework

The pixel-wise labeling of mammograms is not in the common practice of the imaging review process by radiologists. Hence, such annotations, essential for fully supervised methods, are often lacking in mammography datasets. Moreover, regardless of the high prevalence of cancer, the proportion of normal and benign data is still considerably higher, leading to a significant imbalance in the datasets. To this end, we propose a training strategy using negative MG exams, thus requiring an image-wise label only. Such labels are easily recoverable from clinical reports. We introduce a self-supervised approach generating its own synthetic benign and malignant MCs, as presented by Tardy et al.[24] Synthetic lesions are inserted randomly into the image of negative (i.e., normal) breast and an image-wise and pixel-wise ground truth labels are generated. As our framework is fed with small ROIs, we extract patches around those artificial MCs. To balance the training process, we also extract normal patches from without synthesizing.

We guide the learning process with a specific loss for each task. We define $\hat{y}_t$ and $y_t$ respectively as the predictions and ground truths for a task $t$. The categorical cross-entropy ($CCE$) is used as classification loss at each scale $s$ ($\mathcal{L}_{cls_s}$). The $CCE$ formula is defined in the Eq. (1), where $x$ is one of the $K$ classes, in our study $K = 3$: 1) soft tissues, 2) benign MCs, and 3) malignant MCs.

$$CCE = -\sum_{x=1}^{K} y_{cls_x} \ \log(\hat{y}_{cls_x}) \tag{1}$$

For the segmentation loss at each scale $s$ ($\mathcal{L}_{seg_s}$), we used two distinct computation approaches depending on the type of input patch. The first one is applied for patches with pixel-wise ground truths (positive synthetic patches) and is inspired by Milletari et al.[25] The Dice Score Coefficient ($DSC$) is defined in the Eq. (2) where the sums run over the $n$ pixels and $y, \hat{y} \in \mathcal{R}^{H,W}$ with $H$ and $W$ the dimension of the output.

$$DSC = \frac{2\sum_{i=1}^{n} \hat{y}_{seg_i}\ y_{seg_i}}{\sum_{i=1}^{n} \hat{y}_{seg_i}^2\ +\ \sum_{i=1}^{n} y_{seg_i}^2} \tag{2}$$

In the case of soft tissue (negative) patches, we computed the mean absolute error ($MAE$) described in the Eq. (3).

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |\ \hat{y}_{seg_i} - y_{seg_i}| \tag{3}$$

For each task loss, we averaged on all output scales denoted as $\mathcal{S}$ and illustrated in Fig. 1. The global framework loss ($\mathcal{L}_{global}$) is the mean between the average classification and segmentation losses at each level as defined in the Eq. (4).

$$\mathcal{L}_{global} = \frac{1}{2}(\frac{1}{\mathcal{S}}\sum_{s=1}^{\mathcal{S}} \mathcal{L}_{seg_s} + \frac{1}{\mathcal{S}}\sum_{s=1}^{\mathcal{S}} \mathcal{L}_{cls_s}) \tag{4}$$

## 4. EXPERIMENTS

### 4.1 Data

Our framework was trained with an in-house dataset composed of MG exams. Each exam has a malignancy classification according to an American College of Radiology (ACR) risk score guidelines.[26] In this work, we used only normal samples (i.e., $ACR = 1$). To simulate the calcified regions, synthetic lesions were randomly blended on the full-resolution mammography images: a synthetic malignant cluster of MCs is illustrated in Fig. 1. We generated a set of 23,000 unique patches balanced between the three classes: 1) soft tissue, 2) benign MC, and 3) malignant MC. Calcification patches were randomly extracted around the area of blending, with a threshold of 80% lesion pixels required. Normal patches having at least 20% of breast pixels were finally generated randomly from negative MG inputs.

For validation and testing purposes, we used two public datasets: INBreast[14] and Breast MicroCalcifications Dataset (BMCD).[15] Both datasets have pixel-wise annotations from expert radiologists. INBreast contains various types of findings (i.e., isolated calcifications, MC clusters, masses, and architectural distortions). For this study, we used MC cluster cases as malignant and cases with isolated calcifications as benign. BMCD contains annotations only for benign and malignant calcifications. We have therefore retained only those cases where a biopsy revealed a malignant tumor ('MALIGNANT' or 'DCIS' labels only) (more details on labels are available in the original paper [15]).

We extracted a set of 3-class validation and test patches from both datasets with the same strategy. A first extraction round was applied to get patches centered on malignant MC ground truths, maximizing the context of the patch. Then, a linear projection from full-resolution MGs into $512 \times 512$ non-overlapped patches was done, followed by a sorting phase. Patches with benign calcifications were separated from soft tissues analyzing pixel-wise ground truths.

The Tab. 1 summarizes the number of patches generated and used in the three stages of experiments: training, validation, and testing. Approximately 23,000 training, 2,364 validation, and 1,826 test patches have been used during this study.

| | | Normal | Benign | Malignant | **Total** |
|---|---|---|---|---|---|
| Training | Private | 7666 | 7666 | 7666 | **23000** |
| Validation | INBreast[14] | 1030 | 1300 | 34 | **2364** |
| Test | BMCD[15] | 1417 | 383 | 26 | **1826** |

Table 1. Description of training, validation, and test patches generated for this study. Training patches are extracted from a private dataset, while validation and testing phases are realized on two public datasets.

### 4.2 Experimental setup

In this section, we describe the implementation details applied to carry out our experiments and compare our proposed deep CNN with state-of-the-art CNN architectures. We identified three popular CNN architectures adapted for medical imaging classification or segmentation: Resnet-22,[12] ConvNeXt,[13] and UNet3+.[7] In addition, we performed an ablation study on the impact of multi-tasking and multi-scaling outputs. From now on, we define the following terms: mono-scale (MoS) vs. multi-scale (MuS) and mono-task (MoT) vs. multi-task (MuT). Our combination with MuT and MoS (MuTMoS-UNet3+) is similar to the method proposed by Ouyang et al.[19]

As our UNet3+ has approximately 1.4M trainable parameters, we implemented the optimized version of the ResNet-22 proposed by Walsh et al.[27] The authors use separable convolutions instead of classic ones reducing the number of parameters from 2.9M to 460K. Similarly, for a fair comparison to the ConvNext architecture with regard to the number of trainable parameters, we also declined the ConvNeXt-Atto (ConvNeXt-A)[13] in two versions: Zepto (ConvNeXt-Z) and Yocto (ConvNeXt-Y) with 20% and 40% fewer filters respectively. ConvNext-A, Z, and Y have approximately 3.4M, 2.2M, and 1.2M trainable parameters.

To run the experiments, we randomly initialized the model's weights and trained from scratch using the set of 23,000 patches described above. Experiments lasted 500 epochs with a batch size of 3 and a learning rate of $10^{-4}$. Finally, to prevent overfitting, data augmentation techniques were applied. First, the images were randomly flipped with vertical, horizontal, or transpose flips. Second, intensities augmentations were applied amongst the following: gaussian blur, gaussian noise, gamma contrast, and intensity inversion.

For neural network implementation, we used Tensorflow 2.10 library. For image processing operations we used cv2, numpy and scikit-image libraries.

## 5. RESULTS

### 5.1 Classification performances

In this section, we present and discuss classification performances of the proposed method. We compared the following state-of-the-art (SOTA) methods: ResNet-22,[12] ResNet-22 optimized,[27] ConvNeXt-A,[13] ConvNeXt-Z, ConvNeXt-Y as presented above. Secondly, we realized an ablation study for the comparison of MuT vs. MoT and MuS vs. MoS.

We computed two different Area Under the ROC Curve (AUC) scores. The first one ($AUC_{MCs}$) focuses on the classification of calcified areas (benign and malignant) vs. soft tissues, while the second ($AUC_{mlg}$) evaluates the classification of MCs characterized as malignant vs. the rest, i.e., benign calcifications and soft tissues.

The performances of our method and SOTA architectures are presented in the Tab. 2. First, it shows that the traditional ResNet-22[12] performs the worse compared to other methods, and in particular its optimized version. The ResNet-22 optimized challenges with ConvNeXt architectures which achieved similar performances. However, we observe an opposite trend, where the lightened version (ConvNeXt-Y) performs less well than heavier flavors. Overall, SOTA architectures perform better for the characterization of malignant MCs than for the distinguishing classification of calcification areas vs. soft tissues on the validation dataset while the opposite is true for the test dataset. Finally, we see in the Tab. 2 that the proposed MuTMuS-UNet3+ outperforms the SOTA architectures in all but one task. That is, in the task of classification of calcified vs. soft tissues it comes second best with an $AUC_{MCs} = 0.94$, leaving the first place to ConvNeXt-Z having $AUC_{MCs}0.95$.

| | INBreast (validation) | | BMCD (test) | |
|---|---|---|---|---|
| | $AUC_{MCs}$ | $AUC_{mlg}$ | $AUC_{MCs}$ | $AUC_{mlg}$ |
| ResNet-22 | 0.73 | 0.89 | 0.79 | 0.80 |
| ResNet-22 optimized | 0.88 | 0.93 | 0.94 | 0.90 |
| ConvNeXt-A | 0.87 | 0.93 | 0.94 | 0.90 |
| ConvNeXt-Z | 0.89 | 0.91 | **0.95** | 0.88 |
| ConvNeXt-Y | 0.84 | 0.91 | 0.89 | 0.88 |
| MuTMuS-UNet3+ | **0.90** | **0.94** | 0.94 | **0.93** |

Table 2. Comparison of AUC scores achieved by our method and various versions SOTA CNN architectures implemented in the study. Two AUC scores were computed for both validation and test datasets, $AUC_{MCS}$ and $AUC_{mlg}$. Best AUC scores are highlighted.

Then, we conducted the ablation study and reported AUC scores in the Tab. 3. The table shows that the combination of multi-task and multi-scale approaches is beneficial to the framework as it achieved the highest AUC scores on both datasets. Having a multi-scale approach with a single classification task implementation on a UNet3+ is the worst combination, due to its particular design for semantic segmentation tasks. On the other hand, the combination of classification and segmentation outputs on one scale only is not optimal for catching MCs of different sizes. That is, it prevents each scale from detecting objects according to its proper receptive field.

| | | INBreast (validation) | | BMCD (test) | |
|---|---|---|---|---|---|
| Multi-task | Multi-scale | $AUC_{MCs}$ | $AUC_{mlg}$ | $AUC_{MCs}$ | $AUC_{mlg}$ |
| yes | yes | **0.90** | **0.94** | **0.94** | **0.93** |
| no | yes | 0.84 | 0.91 | 0.92 | 0.91 |
| yes | no | 0.79 | 0.77 | 0.87 | 0.87 |

Table 3. AUC scores comparison table for the ablation study on the framework multi-task and multi-scale. Two AUC scores were computed for both validation and test datasets, $AUC_{MCS}$ and $AUC_{mlg}$. Best AUC scores are highlighted.

We note that ResNet-22 optimized, ConvNeXt-A, and ConvNeXt-Z designed for the classification task performed similarly to our network. Nevertheless, our proposed framework based on a modified version of the UNet3+ with multi-task and multi-scale output approaches not only outperformed classic UNet3+, but also the SOTA classification CNNs. It shows the importance of catching multi-scale objects as the MC detection task requires global context and fine-grained details.

### 5.2 Segmentation and activations maps analysis

To allow more complete understanding of the performances of the proposed method, we evaluated its segmentation performances. To this end, we explored the capacities of CNN classifiers to segment the relevant areas. To do so, we generated Gradient-weighted Class Activation Mapping (Grad-CAM) based on Selvaraju et al. method.[28] We aim to obtain a visualized explanation of activations generated by the ResNet-22 optimized and the ConvNext-A and make their decision-making interpretable. We also generated the segmentation predictions of our proposed framework. All predictions were then binarized using the automatic threshold technique proposed by Li et al. method.[29] We purposely excluded other UNet3+ architectures (i.e., MoTMuS and MuTMoS), as performing badly on the classification task (see Tab. 3).

We computed three different semantic segmentation metrics: Recall ($Rec$), Precision ($Prec$), and Dice Score Coefficient ($DSC$) defined in the Eq. (5), (6), and (7) respectively. The Recall quantifies the number of relevant items retrieved, while the Precision quantifies the retrieved elements' relevance. Finally, the Dice coefficient score is a similarity statistic indicator between the predictions and the ground truth.

$$Rec = \frac{TP}{TP + FN} \tag{5}$$

$$Prec = \frac{TP}{TP + FP} \tag{6}$$

$$DSC = \frac{2 \times Rec \times Prec}{Rec + Prec} \tag{7}$$

where TP, FP, and FN denote pixel-wise true positives, false positives, and false negatives respectively.

The metrics are reported in the Tab. 4. Overall, low $DSC$ values are achieved on both datasets for benign and malignant MCs. It can be explained by the coarseness of annotations, as shown by the two examples illustrated in the Fig. 2. Well-contoured annotations are not given for benign calcifications as only some pixels (sometimes only 1 pixel) are labeled, mainly due to the quantity of isolated calcifications to annotate. For malignant annotations, rough delineations are given for both datasets, which is even more amplified on the BMCD dataset (cf. Fig. 2). To that end, we slightly dilated benign calcifications with a $10 \times 10$ squared kernel, making results more interpretable.

Based on the reported metrics, our framework outperformed SOTA methods for the benign MC segmentation task, with the best $DSC$, $Rec$, and $Prec$ performances obtained on both the validation and test datasets. An example of the generated predictions is illustrated in the Fig. 2, A-D. The classifiers tend to generate noisier activations compared to our framework's segmentation which precisely localized the two benign calcifications. High recall rates are obtained by our framework: **0.57** and **0.68** which signifies that low false negatives are generated while having a high true positive prevalence. Meanwhile, the precision score is difficult to analyze due to the small size of annotations. That is, the network generally tends to the over-segmentation.

For the malignant MC's segmentation task, ResNet-22 and our framework both performed well and challenged themselves on validation and test datasets. Our framework achieved the best performances on the validation dataset regarding all metrics, with a Dice of **0.46**, a recall of **0.74**, and a precision of **0.46**. Meanwhile, ResNet-22 achieved the best performances on tests with the highest Dice **0.49**, recall **0.40**, and precision **0.70** with the last one similar to our framework (**0.70**). The results are consistent with the ground truth discussion above, and therefore explainable. Indeed, BMCD ground truths are less precise than INBreast producing many false negatives in annotated areas where not all labelled pixels are truly positives (i.e., depicting calcifications).

| | INBreast (validation) / BMCD (test) | | | | | |
|---|---|---|---|---|---|---|
| | $DSC_B$ | $Rec_B$ | $Prec_B$ | $DSC_M$ | $Rec_M$ | $Prec_M$ |
| ResNet-22 opt. | 0.03 / 0.04 | 0.46 / 0.60 | 0.02 / 0.02 | 0.36 / **0.49** | **0.74** / **0.40** | 0.36 / **0.70** |
| ConvNeXt-A | 0.02 / 0.04 | 0.26 / 0.37 | 0.02 / 0.02 | 0.18 / 0.33 | 0.32 / 0.38 | 0.19 / 0.33 |
| Our framework | **0.17** / **0.25** | **0.57** / **0.68** | **0.13** / **0.18** | **0.46** / 0.36 | **0.74** / 0.27 | **0.46** / **0.70** |

Table 4. Semantic segmentation metrics computed on Grad-CAM of classifiers and segmentation predictions or the proposed method. Metrics were computed on both validation and test datasets for benign and malignant MCs respectively denoted as $B$ and $M$. The best metrics are highlighted.

## 6. DISCUSSIONS AND CONCLUSION

Depicting microcalcifications in mammograms is crucial in the early detection of breast cancer as it may be a sign of malignancy. However, the task remains challenging due to their tiny size, and diversity of shapes essential to characterizing malignancy.

In this study, we propose a novel framework based on a deep CNN architecture designed for the specific problem of MC detection and classification. The framework is designed to distinguish soft tissues from benign and malignant calcified tissues. To achieve this task, we rely on a two-channel segmentation and three-class classification outputs strategy. Underneath it fuses multi-scale features improving overall performances (see Tab. 3). Moreover, it requires only limited annotations thanks to the use of synthesized data.

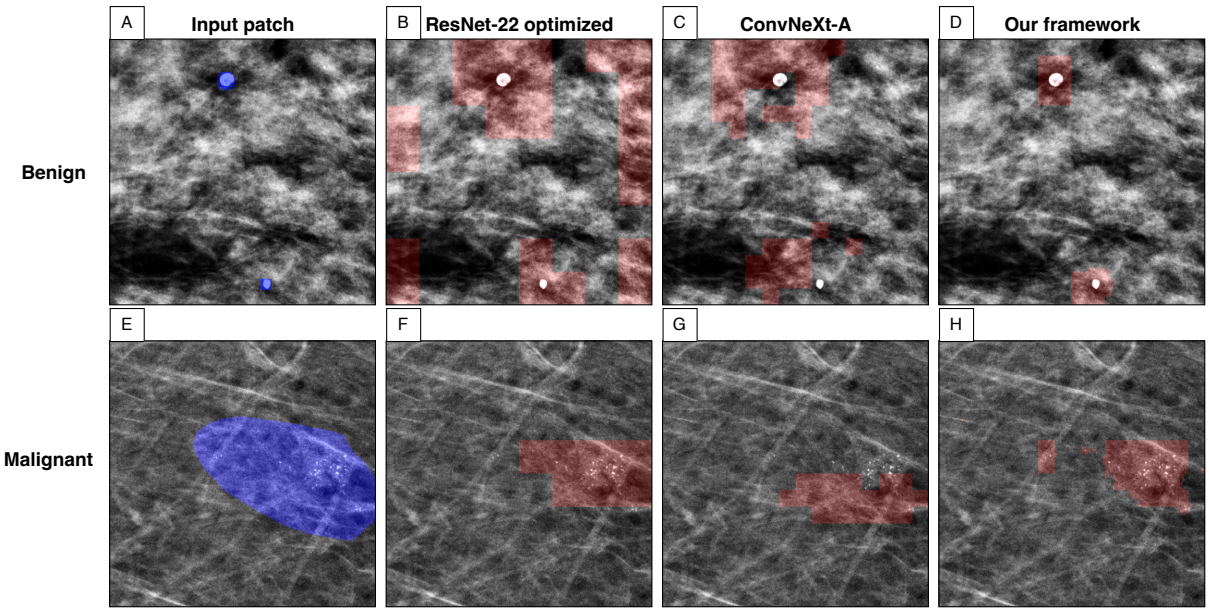


Figure 2. From left to right: (A, E) input patch with ground truth in blue; (B, F) activations heatmaps of ResNet-22 optimized; (C, G) activations of ConvNeXt-A; (D, H) semantic segmentation output of MuTMuS-UNet3+. Top row (A-D): Benign MCs patch, and Bottom row (E-G): Malignant MCs patch. Blue and red areas denote ground truths and predictions respectively.

We compare the performances of our method to several flavors of popular state-of-the-art CNN networks, such as ResNet and ConvNeXt. We also perform an ablation study highlighting the relevance of the proposed multi-output and multi-scale aspects. We use two popular public datasets for the validation and test: INBreast and Breast MicroCalcifications Dataset (BMCD). To the best of our knowledge, we are the first to claim semantic segmentation metrics on the second dataset which is quite recent in the community. Both classification and segmentation performances are explored and discussed in a patch-wise manner.

We observe that our framework outperforms the state-of-the-art methods in almost all the tasks. The highest improvement is achieved on BMCD dataset on the task of binary classification of malignant MCs vs. other tissues (soft and benign calcifications) with a value of **AUC=0.93**. In addition, our framework generates more precise semantic segmentation on both channels, in particular with the pixel-wise precision of **0.70** on BMCD.

We want to draw the reader's attention to an eventual difficulty in the interpretation of the segmentation metrics due to their approximate nature (see Fig. 2 E.). Hence, the reported DICE scores may appear low compared to what can usually be found in the literature. While we observe higher precision of the proposed method, the interpretability of the generated output should be studied further.

In this work, we focused on the patch-wise classification and segmentation analysis. We noted promising results and consistent improvement compared to the state-of-the-art method. However, to bring a more clinically relevant guidance, an image-wise approach is expected. We aim to address it in the future.

## ACKNOWLEDGMENTS

This research is supported by the CIFRE program granted by the French ANRT organism under contract no. 2022/155.

## REFERENCES

[1] Siegel, R. L., Miller, K. D., Wagle, N. S., and Jemal, A., "Cancer statistics, 2023," *CA: A Cancer Journal for Clinicians* **73**, 17–48 (jan 2023).

[2] Coleman, C., "Early Detection and Screening for Breast Cancer," *Seminars in Oncology Nursing* **33**, 141–155 (may 2017).

[3] O'Grady, S. and Morgan, M. P., "Microcalcifications in breast cancer: From pathophysiology to diagnosis and prognosis," *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer* **1869**, 310–320 (apr 2018).

[4] Tot, T., Gere, M., Hofmeyer, S., Bauer, A., and Pellas, U., "The clinical value of detecting microcalcifications on a mammogram," *Seminars in Cancer Biology* **72**, 165–174 (jul 2021).

[5] Le, E. P., Wang, Y., Huang, Y., Hickman, S., and Gilbert, F. J., "Artificial intelligence in breast imaging," *Clinical Radiology* **74**, 357–366 (may 2019).

[6] Badawy, E., Shalaby, F. S., Saif-El-Nasr, S. I., Elyamany, A. M., Mohamed, R., and Hegazy, A., "The synergy between AI and radiologist in advancing digital mammography: comparative study between standalone radiologist and concurrent use of artificial intelligence in BIRADS 4 and 5 female patients," *Egyptian Journal of Radiology and Nuclear Medicine 2023 54:1* **54**, 1–12 (nov 2023).

[7] Huang, H., Lin, L., Tong, R., Hu, H., Zhang, Q., Iwamoto, Y., Han, X., Chen, Y. W., and Wu, J., "UNet 3+: A Full-Scale Connected UNet for Medical Image Segmentation," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings* **2020-May**, 1055–1059 (apr 2020).

[8] Zhou, Q., Yang, W., Gao, G., Ou, W., Lu, H., Chen, J., and Latecki, L. J., "Multi-scale deep context convolutional neural networks for semantic segmentation," *World Wide Web* **22**, 555–570 (mar 2019).

[9] Jiang, Y., Liu, W., Wu, C., and Yao, H., "Multi-Scale and Multi-Branch Convolutional Neural Network for Retinal Image Segmentation," *Symmetry 2021, Vol. 13, Page 365* **13**, 365 (feb 2021).

[10] Zhou, Y., Chen, H., Li, Y., Liu, Q., Xu, X., Wang, S., Yap, P. T., and Shen, D., "Multi-task learning for segmentation and classification of tumors in 3D automated breast ultrasound images," *Medical Image Analysis* **70**, 101918 (may 2021).

[11] Zhang, C. and Zhang, Z., "Improving multiview face detection with multi-task deep convolutional neural networks," *2014 IEEE Winter Conference on Applications of Computer Vision, WACV 2014* , 1036–1041 (2014).

[12] 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., and Geras, K. J., "Deep Neural Networks Improve Radiologists' Performance in Breast Cancer Screening," *IEEE Transactions on Medical Imaging* **39**, 1184–1194 (apr 2020).

[13] Woo, S., Debnath, S., Hu, R., Chen, X., Liu, Z., Kweon, I. S., Xie, S., and Ai, M., "ConvNeXt V2: Co-Designing and Scaling ConvNets With Masked Autoencoders," (2023).

[14] Moreira, I. C., Amaral, I., Domingues, I., Cardoso, A., Cardoso, M. J., and Cardoso, J. S., "INbreast: Toward a Full-field Digital Mammographic Database," *Academic Radiology* **19**, 236–248 (feb 2012).

[15] Loizidou, K., Skouroumouni, G., Pitris, C., and Nikolaou, C., "Digital subtraction of temporally sequential mammograms for improved detection and classification of microcalcifications," *European Radiology Experimental* **5**, 1–12 (dec 2021).

[16] Sakaida, M., Yoshimura, T., Tang, M., Ichikawa, S., and Sugimori, H., "Development of a Mammography Calcification Detection Algorithm Using Deep Learning with Resolution-Preserved Image Patch Division," *Algorithms 2023, Vol. 16, Page 483* **16**, 483 (oct 2023).

[17] Quintana, G. I., Li, Z., Vancamberg, L., Mougeot, M., Desolneux, A., and Muller, S., "Exploiting Patch Sizes and Resolutions for Multi-Scale Deep Learning in Mammogram Image Classification," *Bioengineering 2023, Vol. 10, Page 534* **10**, 534 (apr 2023).

[18] Shen, L., Margolies, L. R., Rothstein, J. H., Fluder, E., McBride, R., and Sieh, W., "Deep Learning to Improve Breast Cancer Detection on Screening Mammography," *Scientific Reports* **9** (dec 2019).

[19] Ouyang, X., Che, J., Chen, Q., Li, Z., Zhan, Y., Xue, Z., Wang, Q., Cheng, J. Z., and Shen, D., "Self-adversarial Learning for Detection of Clustered Microcalcifications in Mammograms," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **12907 LNCS**, 78–87 (2021).

[20] Zhang, F., Luo, L., Sun, X., Zhou, Z., Li, X., Yu, Y., and Wang, Y., "Cascaded Generative and Discriminative Learning for Microcalcification Detection in Breast Mammograms," (2019).

[21] Gerbasi, A., Clementi, G., Corsi, F., Albasini, S., Malovini, A., Quaglini, S., and Bellazzi, R., "DeepMiCa: Automatic segmentation and classification of breast MIcroCAlcifications from mammograms," *Computer Methods and Programs in Biomedicine* **235**, 107483 (jun 2023).

[22] Feng, S., Zhao, H., Shi, F., Cheng, X., Wang, M., Ma, Y., Xiang, D., Zhu, W., and Chen, X., "Cpfnet: Context pyramid fusion network for medical image segmentation," *IEEE Transactions on Medical Imaging* **39**, 3008–3018 (oct 2020).

[23] Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. L., "DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **40**, 834–848 (apr 2018).

[24] Tardy, M. and Mateus, D., "Looking for abnormalities in mammograms with self-and weakly supervised reconstruction," *IEEE Transactions on Medical Imaging* **PP**, 1–1 (jan 2021).

[25] Milletari, F., Navab, N., and Ahmadi, S. A., "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," *Proceedings - 2016 4th International Conference on 3D Vision, 3DV 2016* , 565–571 (dec 2016).

[26] D'Orsi, C. J., [*2013 ACR BI-RADS Atlas: Breast Imaging Reporting and Data System - Acr*], American College of Radiology (2014).

[27] Walsh, R. and Tardy, M., "A Comparison of Techniques for Class Imbalance in Deep Learning Classification of Breast Cancer," *Diagnostics* **13** (nov 2023).

[28] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D., "Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization," (2017).

[29] Li, C. H. and Lee, C. K., "Minimum cross entropy thresholding," *Pattern Recognition* **26**, 617–625 (apr 1993).