DeepLIMa: Deep Learning Based Lesion Identification in Mammograms

Zhenjie Cao[1], Zhicheng Yang[1], Xiaoyan Zhuo[1,2,*], Ruei-Sung Lin[1], Shibin Wu[3],
Lingyun Huang[3], Mei Han[1], Yanbo Zhang[1,†], Jie Ma[4]
[1]PingAn Tech, US Research Lab, USA
[2]University of Massachusetts, Lowell, USA
[3]Ping An Technology, China
[4]Shenzhen People's Hospital, China
[†]Corresponding author: yanbozhang007@gmail.com

Abstract

Mammography is a major technique for early detection of breast cancer, typically through detection of masses or calcifications. However, how to help radiologists efficiently recognize these lesions remains a challenging problem. In this paper, we propose comprehensive deep learning based solutions to respectively detect masses and segment calcifications in mammograms. To achieve the optimal mass detection performance, our method combines Faster R-CNN with Feature Pyramid Networks, Focal Loss, and Non-Local Neural Networks. We thoroughly compare the proposed method and competing methods on three public datasets and an in-house dataset. The best detection results on our in-house dataset are an average precision of 0.933 and a recall of 0.976. Regarding calcification segmentation, we design a series of pre-processing methods including window adjustment, breast region extraction and artifact removal to normalize mammograms. A U-Net model with group normalization is then applied to segment calcifications. The proposed method is validated on our in-house dataset using a newly designed evaluation metric. The experimental results have demonstrated the great potential for this task.

1. Introduction

Medical imaging has been a revolutionary way for medical professionals to diagnose and treat medical diseases over the past decades. However, interpretation of medical images is a complex task, which can only be performed by medical professionals who have been extensively trained on reading medical images and have long-time clinical experience. With the recent advance of Artificial Intelligence (AI) with the emphasis of *deep learning* methods, numerous Convolutional Neural Networks (CNN) based approaches for computer-vision-related applications have been proposed [14, 19]. Many research projects also apply CNN-based methods to Computer-Aided Detection (CADe) and Computer-Aided Diagnosis (CADx) of medical images for effectively assisting doctors (or even automatically) to locate lesions and determine if they are benign, or malignant (commonly known as *cancer*) in medical images [17]. This promising approach is expected to reduce radiologists' workload and to accelerate the diagnosis process while improving the diagnosis accuracy. These intelligent applications are also capable of integrating with the e-health system to generate a comprehensive clinical report, and/or to provide potential personal assistance, etc. [13].

Among these applications on medical images, detecting lesions in *mammography*, the primary imaging technique used for breast screening process, is gaining the increasing attention during recent years [23], because breast cancer is the leading type of cancer in woman, accounting for massive death worldwide every year [4]. There are mainly four types of abnormality patterns in mammograms: mass, calcification, asymmetry, and architectural distortion. In particular, *mass* and *calcification* account for the vast majority of all breast cancer findings. Compared to calcification, mass does not have a clear outline, and looks similar to the normal tissues. Due to their own distinct characteristics, developing one universal solution to both of them is not ideal. To better understand that, we list the existing challenges and our solutions as follows.

Challenge 1: *Detection problem vs. Segmentation problem.* As aforementioned, one single solution to tackle both two types of lesions is not desired. How to formalize them as either segmentation or detection problem is critical.

Solution: According to the common computer vision guidance [5], mass, which is always viewed as a *region* and has a vague outline, is more suitable for detection problem. On the other hand, formalizing calcification identification

[*] This work was done during Xiaoyan Zhuo's internship at PingAn Tech, US Research Lab, USA.

as a segmentation problem is an appropriate way thanks to the clear contour of calcification. In our solution, we exploit an enhanced version of Region proposals with CNNs (R-CNN) [6], Faster R-CNN [22] to solve the mass detection problem, while a U-Net [24] structure is leveraged to segment calcification. We will detail our approaches and contributions later in this section.

Challenge 2: *Inconsistency of calcification annotation.* Calcifications have various shapes and distributions. Compared to the satisfactory consistency of mass annotation, calcification annotation depends on each radiologist's personal annotation preference very much. Two examples with the different annotations by radiologists can be found in Fig. 7 later, who labeled every single calcification or assumed the calcifications a large region. Such difference results in an inconsistent and imperfect dataset that limits the model's learning capacity.

Solution: We improve the annotation consistency in the calcification dataset and to develop efficient deep learning networks without introducing any additional overhead into the entire framework.

We here briefly review the related research works. Regarding mass detection, authors in [28] leveraged conventional machine learning methods in K-means and support vector machine (SVM) classification. Abstract patterns were extracted from malignant and benign masses, and were then used for the prediction model training. Recently, various deep-learning-based frameworks substantially outperform traditional machine learning methods for object detection tasks. The Faster R-CNN improved the overall detection performance and significantly reduced the processing overhead [22]. A recent work applied the Faster R-CNN framework on the large mixed public mammogram datasets to detect masses on mammograms [23]. For calcification segmentation, authors in [7] leveraged a set of preprocessing schemes with a simplified pulse-coupled neural network to detect micro-calcification clusters. A U-net based reconstruction framework was introduced in [27] to extract the micro-calcification proposals.

In this paper, we present our comprehensive deep-learning-based solution for mass detection and calcification segmentation tasks. Our key contributions are summarized below:

1. We adopt three effective neural network techniques, namely Feature Pyramid Networks (FPN) [15], Focal Loss [16] and Non-Local Neural Networks [25], and integrate them within the Faster R-CNN framework for mass detection. The three modules employed all results in performance improvement, and the proposed method outperforms the baseline methods by a large margin.

2. We propose a modified U-Net framework with a novel series of pre-processing methods to mitigate the impact of imperfect images and inconsistent annotations in calcification datasets. To address the issues raised by the imperfect data and labels, we develop a series of data and label processing methods, including artifact reduction, label classification and consistency processing. With a newly designed evaluation metric, a thorough evaluation is done overall and each category of calcifications.

3. Aside from using three public datasets, we constitute our in-house dataset. The proposed methods and models are applied on our private mass and calcification datasets, and achieve the outstanding results. Such demonstration empowers the future delivery of our approaches to the collaborative hospitals for practical on-site use.

The rest of this paper is organized as follows. Detailed descriptions of the proposed method are given in Sec. 2. Sec. 3 introduce our public dataset and in-house dataset. We evaluate our proposed methods in Sec. 4. Discussion and future work are presented in Sec. 5. We conclude this paper in Sec. 6.

2. Methods

In this section, we describe the details of our methods of mass detection and calcification segmentation.

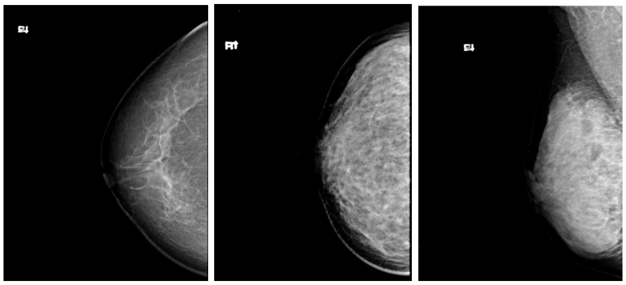
2.1. Mass detection

2.1.1 Faster R-CNN

Faster R-CNN is a two-stage deep-learning-based object detection model. In the Faster R-CNN, a backbone network is adopted to generate feature maps. Based on feature maps, region candidates are first automatically generated by a region proposal networks (RPN) instead of the previous slow selective search algorithm in the R-CNN version [6], and then a CNN-based network is used to classify the object class and detect the bounding box. Moreover, the whole backbone convolution layers are shared layers which are used for both region proposal networks and classification head. This way, the overhead of the entire framework is significantly reduced, compared to other previous transformed versions of R-CNN.

2.1.2 FPN

FPN is a multi-scale algorithm for object detection proposed in [15]. Without introducing additional calculation overhead, FPN exploits the inherent hierarchical architecture in the process of generating the top feature map layer in the backbone network, and efficiently extracts multi-scale



(a) Least dense (b) Medium dense (c) Densest

Figure 1: Mammograms with different breast densities.

feature maps to constitute a pyramid structure for the following prediction tasks.

Since a typical mass does not have distinct outlines, a clear breast background is expected where the mass can be easily identified. However, due to the variation of radiation dose used for mammography, a clear background is not always achieved. As mammogram images have complex structures, identifying mass lesions from a dense background requires specialists to observe more areas and to distinguish the possible lesions from the normal dense breast texture. On the other hand, if a mass lies within a clear background, a smaller and more local receptive field is possibly adequate to locate it. FPN is a practical approach to facilitate this scalable detection, achieved by the fusion of multi-level feature maps with different sizes.

2.1.3 Focal Loss

As aforementioned in Sec. 2.1.2, detecting lesions in dense breasts is a challenging task, because the contrast of background will heavily impact the model's performance. Fig. 1 presents breast examples with three different densities. In these examples, Fig. 1c can be regarded as one type of "hard examples" in the lesion detection task, as it has the densest breast among these three examples. This observation enables us to reconsider the importance of "easy examples" (masses with clear backgrounds) and "hard examples" (masses with bright backgrounds) during the training procedure. After the several beginning epochs, the easy examples tend to have less contribution to the loss. Instead, the contribution of hard examples to the loss should be increased. As observing that the original Faster R-CNN usually predicts many false positives for mass detection, we replace the original loss function in Faster R-CNN with the state-of-the-art focal loss, which utilizes a weighting strategy and focuses on hard examples. The focal loss is defined as follows [16]:

$$FL(\not{p}) = -(1-\not{p})^{\gamma}\log(\not{p}), \tag{1}$$

where

$$\not{p} = \begin{cases} p & \text{if } y = 1 \\ 1-p & \text{otherwise,} \end{cases} \tag{2}$$

where $y \in \{-1, +1\}$ refers to the ground-truth binary class; $p \in [0, 1]$ denotes the predicted probability of the class with label $y = 1$; γ is a focusing variable, which is not less than 0.

2.1.4 Non-Local Neural Networks

The *Non-local Means* (NL-means) is a traditional computer vision algorithm originally used for image denoising [1]. Instead of calculating the mean value of a target pixel's "local" surrounding pixels, this algorithm computes the mean of all "non-local" pixels of the entire image, weighted by the similarity of each pixel to the target pixel. Therefore, the global details that might be ignored by the local mean approach can be maintained. Recently, authors in [25] proposed non-local neural networks that applied the strategy of NL-means to the modern deep learning architecture and has demonstrated its efficacy of capturing long-range dependencies. Such dependency is also an important concern in mammograms. The original receptive field in the backbone network of Faster R-CNN model might not consider sufficient global information to calculate the response of a single position.

The key formula of the non-local neural networks [25] is:

$$\mathbf{v}_i = \frac{1}{C(\mathbf{u})}\sum_{\forall j} f(\mathbf{u}_i, \mathbf{u}_j)g(\mathbf{u}_j), \tag{3}$$

where **u** refers to an image; i is the index of a target position; j denotes the index of all possible positions through the image **u**; f calculates a scalar relationship between i and j; g denotes an operation at the position j; **v** represents the calculated outcome of this equation, which is normalized by a factor $C(\mathbf{u})$. We here adopt the embedded dot product version as the function f.

$$f(\mathbf{u}_i, \mathbf{u}_j) = \theta(\mathbf{u}_i)^T\phi(\mathbf{u}_j), \tag{4}$$

where $\theta(\mathbf{u}_i) = W_\theta \mathbf{u}_i$; $\phi(\mathbf{u}_j) = W_\phi \mathbf{u}_j$; W_θ and W_ϕ denote the two target weight matrices, respectively. To avoid introducing much extra computation cost, we limit the number of non-local blocks in the late stage of our backbone network.

2.2. Calcification Segmentation

2.2.1 Data Pre-processing

Window Adjustment: Mammograms collected from different machines may be stored with different numbers of bits. Hence, it is necessary to adjust the window to ensure that the breast tissues can be well displayed in the same value range. DICOM files contain the window center c and

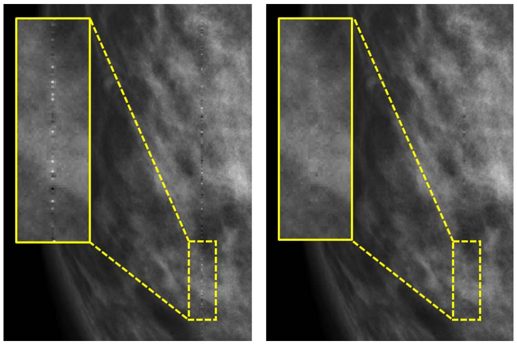


Figure 2: An example of artifacts. Left: original image. Right: corrected image.

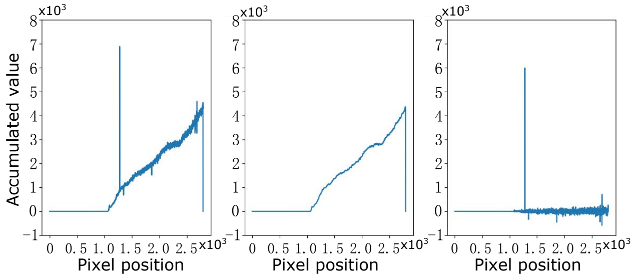


Figure 3: Accumulated profiles for the localization of artifacts. From the left to the right, plots are $h(x)$, $\hat{h}(x)$ and $h(x) - \hat{h}(x)$, respectively.

window width w. We linearly map the pixel values $c - w/2$ and $c + w/2$ as 0 and 255, respectively. In the window adjusted image, the pixels with negative values are set to 0.

Breast Region Extraction: After the window adjustment, pixel values in the background of breast images are zero. However, there exist non-breast components, such as letters. To remove non-breast components, we use morphology methods to obtain the connected regions of non-zero pixels. The breast region is the one with the largest area and is preserved, and values of the rest pixels are set to zero.

Artifact Removal: Mammograms usually contain various artifacts. The left figure in Fig. 2 shows typical artifacts, where some black and bright pixels appear in a line. These bright dots are very similar to micro-calcifications, so it is essential to remove these artifacts before training and testing. This kind of artifact has the following two characteristics: 1) The size of each black or bright artifact dot is about 1 pixel; 2) The distribution of artifact dots is along a vertical line. According to these features, we designed an artifact localization and removal strategy.

Let $\mathbf{u}(x, y)$ be the value of pixel (x, y), and $\Omega(x, y)$ defines a set of neighboring pixels of pixel (x, y). $\mathbf{u}_j$ is the j^{th} neighboring pixel of (x, y) for $j \in \Omega(x, y)$. In this work, 8-neighborhood is selected as $\Omega(x, y)$. We find the minimum absolute value of a pixel and its neighboring pixels. The obtained minimum absolute values are accumulated along the vertical direction to generate function $h(x)$. This operation is summarized in the formula (5).

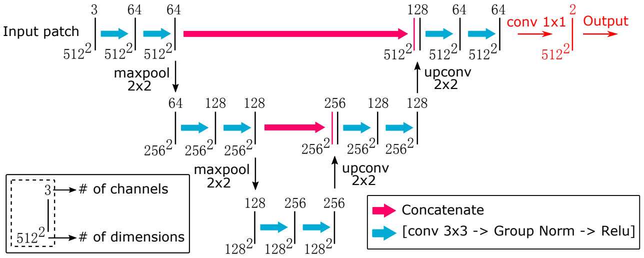


Figure 4: Architecture of U-Net model for calcification segmentation.

$$h(x) = \sum_{y} \min\{|\mathbf{u}(x, y) - \mathbf{u}_j|\}|_{\{j \in \Omega(x,y)\}} \tag{5}$$

The left plot in Fig. 3 presents the $h(x)$ of the example showed in Fig. 2. There is an obvious pulse at $x = 1279$, which matches the location of artifacts. To automatically detect this location, we process $h(x)$ with median filtration to get a smooth curve $\hat{h}(x)$ as shown in the middle plot of Fig. 3. Then $h(x) - \hat{h}(x)$ removes the contribution of background in the image as shown in the right plot in Fig. 3, and a threshold is applied to detect the location of artifacts. Finally, each pixel value $\mathbf{u}(x, y)$ along artifact lines is replaced with median of its neighborhood $\Omega(x, y)$, and artifacts are removed. As shown in the right of Fig. 2, artifacts have been successfully removed while other pixels remain unchanged.

2.2.2 Training Methods

U-Net is a CNN-based segmentation network which has been proved effective in medical images [24]. Fig. 4 illustrates the architecture of U-Net used in this work. This U-Net consists of 3 downsample stages and 3 upsample stages with skip connections. Each stage has two convolution layers, and each one followed by a group normalization [26] and ReLu [21].

Instead of using an entire image as the training input, we crop it into a number of 512×512 patches. The corresponding cropping procedure is also performed on the pixel-level binary mask of this image. Consequently, a patch-based training set is constructed.

Since every patch still has the dominant proportion of background (black) pixels, our patch-based training dataset is extremely imbalanced in terms of positive vs. negative pixels. To overcome this issue, we sample positive and negative patches in every mini-batch up to the ratio of 1:1. Besides, to avoid the parameter-updating leans toward the negative pixels, we apply weighted cross-entropy loss to aggregate more penalty on positive samples. The Adam optimizer [10] is utilized, and the learning rate is initially set as 0.01, and then decreases gradually.

In the test stage, the pixel-based prediction of a patch from the test set is performed. All patches then reconstruct the entire image by their original locations and map their pixel-based prediction probabilities to every pixel of the whole image.

3. Datasets

3.1. Public Dataset

Our public dataset comprises Curated Breast Imaging Subset of Digital Database for Screening Mammography (CBIS-DDSM) [11], INbreast [20], and Breast Cancer Digital repository (BCD) [18]. The original DDSM database [12] is a well-known public breast disease database. However, since the mammograms were collected using different equipment by several hospitals in the United States, the image quality was diverse. We thus adopt CBIS-DDSM, an updated version of DDSM, which eliminates the defective images and forms a standardized subset for more convenient use. CBIS-DDSM has ∼1,600 images for mass lesions and ∼1,630 images for calcification lesions, respectively. INbreast has a total of 115 mass and 128 calcification images. BCD dataset consists of 535 mass images only.

3.2. In-House Dataset

We are collecting an in-house dataset from the collaborative hospitals to evaluate the proposed methods for future on-site delivery. The dataset currently consists of 910 mass images with a total of 1128 mass labels and 1776 calcification images, containing 8157 calcification labels, which were annotated by two radiologists. Fig. 7 presents the calcification examples of our in-house dataset.

3.3. Strategy of Dataset Usage

For our mass detection task, the CBIS-DDSM data are randomly split as training, validation and test data with the percentage of 60%, 20%, and 20%, respectively. The other two public datasets (INbreast and BCD) along with the separated training subset of CBIS-DDSM constitute the training set, while the remaining validation and test subsets of CBIS-DDSM are respectively sole validation and test sets. Compared to the plentiful mass lesion data in the public dataset, the mass samples in our in-house dataset are currently limited. Similar to the apportionment of CBIS-DDSM, the in-house dataset is randomly split as training, validation, and test data with the proportions of 60%, 20%, and 20%, respectively. The best model obtained using the public dataset is fine-tuned on the in-house training data, then the fine-tuned model is validated and evaluated on the in-house validation and test data, respectively.

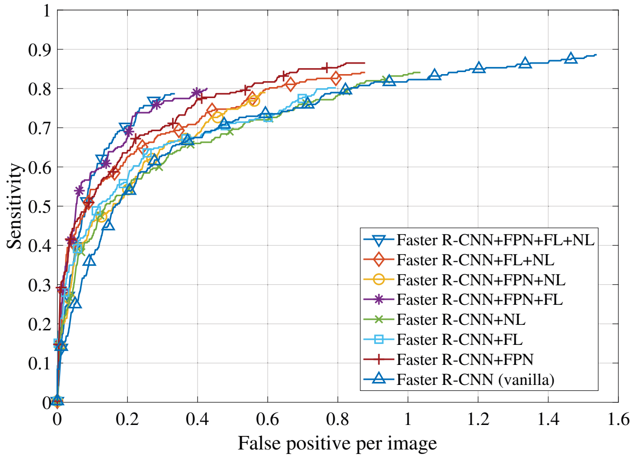


Figure 5: FROC curves of different models. "FL" refers to focal loss, "NL" is non-local operation, and "vanilla" denotes the original Faster R-CNN model.

For calcification segmentation, the public datasets are not adopted to train or evaluate the model. BCD dataset does not contain calcification labels, while INbreast has labels for only part of the calcifications. Although CBIS-DDSM contains complete calcification annotations, its annotation areas are much larger than their actual calcification lesions, which introduces too much unnecessary background. Additionally, its annotation style is very different from our in-house dataset. Hence, we only use the in-house dataset to train and test the calcification segmentation model. The in-house data are randomly grouped as training, validation and test data with proportions of 60%, 20% and 20%, respectively.

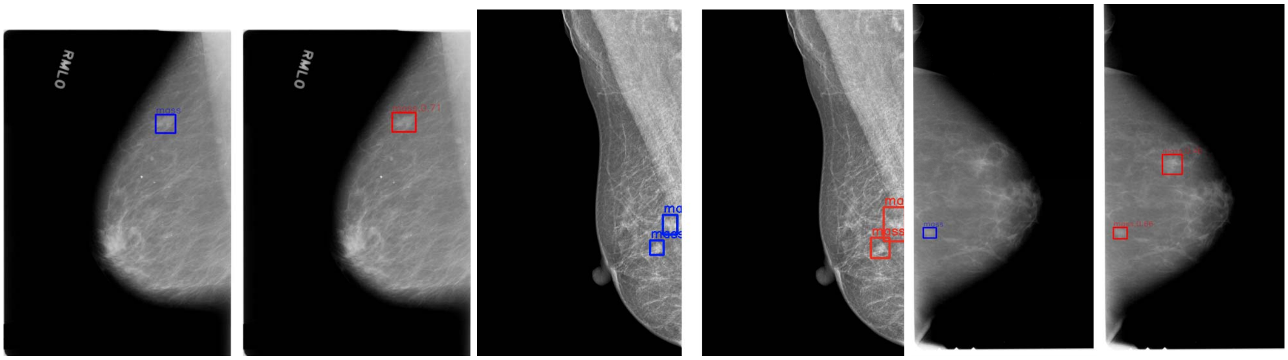
4. Experiments and Results

4.1. Mass Detection

4.1.1 Settings and Parameters

We here describe the settings and the key parameters applied to mass detection using the Faster-RCNN model. ResNet-50 [8] is used as the backbone network of our Faster R-CNN, where the hyper-parameters are loaded from the pre-trained model on ImageNet [3]. Each original training image is down-sampled to a small size to ensure that the short edge has 1200 pixels. The Adam optimization method [10] is used in the model training. There are 500 steps in each epoch, and the training process terminates after 200 epochs.

In the Faster R-CNN, the amount of proposals that are sent to the Fast R-CNN classifier has an impact on both the computation resource and calculation workload for the head function. Since there could only be a limited number of mass lesions in one mammogram, we set a small number of



(a) One mass is successfully predicted.

(b) Multiple masses are successfully predicted.

(c) One prediction is correct while one false positive occurs.

Figure 6: Example prediction results of mass detection in mammograms. Bounding boxes in ground truth are labeled in blue, and the predicted masses are highlighted in red bounding boxes, respectively.

Table 1: Prediction results of different models on the public dataset. "FL" refers to focal loss, "NL" is non-local, and "vanilla" denotes the original Faster R-CNN model.

Method	AP	Recall
Faster R-CNN (vanilla)	0.762	0.950
Faster R-CNN+FPN	0.789	0.965
Faster R-CNN+FPN+FL+NL	**0.805**	**0.977**

"candidate" proposals generated from RPN to the classification head. Since the amount of training images is not very large, there is a risk of over-fitting issue. We utilize several data augmentation methods, such as horizontally flip, translation, and scaling, in training to generate more data.

4.1.2 Effectiveness of Modules

We first evaluate the individual and joint effectiveness of the three described modules embedded into the Faster R-CNN model. The commonly used paradigm, Free-response Receiver Operating Characteristic (FROC), is adopted as our metric. Fig. 5 shows that the FROC curves of various combinations of the modules with the vanilla Faster R-CNN model. We observe that 1) the Faster R-CNN model with all of three modules together outperform the others; 2) the effectiveness of FPN is the most outstanding; 3) the other two modules, focal loss and non-local operation, indeed improve the model. However, compared to their individual efficacy, the joint improvement is notable.

4.1.3 Results

We first present the prediction results of three representative models on our public dataset in Table 1. Average precision (AP) is a widely used evaluation metric in object detection models. It is calculated as mean precision over several recall levels. When the short edge of an image is resized to 1,200 pixels, the Faster R-CNN model with all three modules achieves the best detection results of 0.805 in AP, and 0.977 in recall, respectively. Fig. 6 provides three example prediction results of mass detection in mammograms. As we can see, one or multiple masses are successfully located (Fig. 6a,6b). One false positive shown in Fig. 6c occurs. This might be caused by the sharpness of the wrongly predicted area against the background.

We then apply the best model obtained using the public dataset to our in-house dataset, reporting an AP of 0.933 and a recall of 0.976. In addition, the best resultant point in precision-recall (PR) curve is 0.91 (precision) and 0.90 (recall), respectively.

4.2. Calcification Segmentation

4.2.1 Data Preparation

Remind that the inconsistency issue of calcification annotation is described in Sec. 1, and some annotation examples are shown in Fig. 7. As shown in the right sub-figure, there are many crowded calcification dots. It will take too many efforts for radiologists to annotate every single calcification dot. Hence, these crowded calcifications were annotated as large regions. Besides, different from small scale calcifications, some calcifications exist along vessels as shown in the middle of Fig. 7. Because calcifications were annotated in different manners, we further processed the dataset for easy use and evaluation. According to the area and shape of calcification labels, we classified labels into three groups: dots, vessels and large areas. Table 2 summaries the number of each label type.

The maximum size of a mammogram image reaches 3518×2800, which is too large to fit in GPU memory. A sliding window with the size of 512×512 moves in the original images to extract small patches, and the moving step was set to 256. The patches containing only black back-

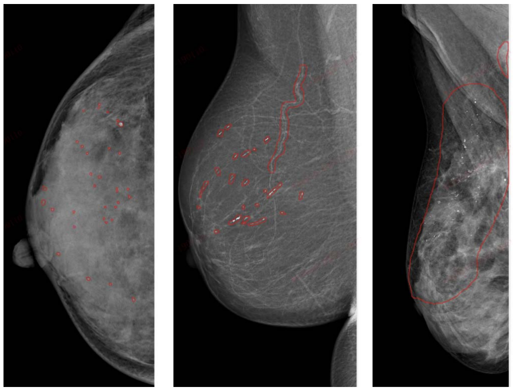


Figure 7: Three examples of calcification labels. Calcification regions were annotated by radiologists with red outlines. From the left to right, these examples indicate types of dots, dots and vessel, and large region, respectively.

Table 2: Numbers of calcification labels.

Total	**Dots**	**Vessels**	**Large regions**
7903	6970	146	787

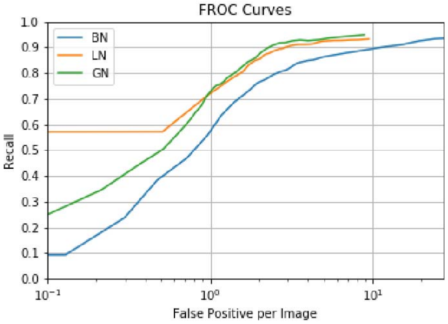


Figure 8: FROCs of the calcification segmentation.

ground were removed, and tens of thousands of patches were obtained as the training and validation data. To improve the label consistency, we adopted two steps to process training data. On the one side, the patches containing large label regions or vessels were not included in the training and validation sets. One the other side, the labels with area fewer than 900 pixels were dilated to 900 pixels.

4.2.2 Evaluation Metric

Some radiologists manually annotated calcification regions are much larger than single calcification dots, while a predicted area usually only covers a single calcification dot. Thus, there usually exist many predicted areas in one annotated region. Because the remarkable different sizes between prediction and annotation, the commonly used *Intersection over Union* (IoU) is not a good criterion for the evaluation of calcification segmentation in this work. Hence, we introduce *Intersection over Prediction* (IoP) as an evaluation metric. IoP is defined as the intersection area between ground truth and prediction divided by the area of prediction. If the IoP of a predicted calcification is greater than a given threshold (0.25 in this work), this prediction is assumed as a true positive. Otherwise, it is considered as false positive. If a ground truth area contains multiple true positive predictions, it only counts as one true positive.

Table 3: Recall at k false positive per image for all methods on the in-house dataset.

Method	$k = \mathbf{1}$	$k = \mathbf{2}$	$k = \mathbf{5}$	$k = \mathbf{10}$
U-Net BN	0.596	0.768	0.862	0.896
U-Net LN	0.724	0.852	0.924	0.932
U-Net GN	**0.737**	**0.869**	**0.932**	**0.944**

Table 4: Recall of three calcification types at k false positive per image for U-Net GN on the in-house dataset.

Method	$k = \mathbf{1}$	$k = \mathbf{2}$	$k = \mathbf{5}$	$k = \mathbf{10}$
Dot	0.691	0.843	0.918	0.933
Vessel	0.971	1.000	1.000	1.000
Large Region	0.912	0.974	0.988	0.991

4.2.3 Results

We here present our results of calcification segmentation on our in-house dataset. A U-Net model with group normalization (GN) was adopted. We also implemented U-Net models with batch normalization (BN) and layer normalization (LN) as competing methods. Fig. 8 compares the FROC curves of U-Net with BN, LN and GN, respectively. It is clear that the performance of U-Net methods with LN and GN are remarkably better than that of U-Net BN. Recall of U-Net GN is slightly higher than U-Net LN when the false positive per image is above 1. Table 3 lists recall at k false positive per image, which confirms the superior performance of U-Net GN to the competing methods. To demonstrate the capability of U-Net GN for a specific type of calcification, recall of three types of calcifications are give in Table 4. The type of vessel has the highest recall among all three, and all vessel calcifications were detected when the false positive per image is no less than 2. Followed by the class of large regions, its recall is higher than 0.91 when false positive per image is no less than 1. In comparison, dots have the lowest recall.

Fig. 9 presents some typical calcification segmentation results. In the top two rows, all the methods have detected calcifications in the large region and vessel. However, U-Net BN missed a few parts of calcifications in the vessel. In the third row, U-Net BN missed two calcification dots, which can be detected by U-Net LN and GN. In the bottom two rows, U-Net BN presents many false positives (red dots), while U-Net GN achieves the most precise results.

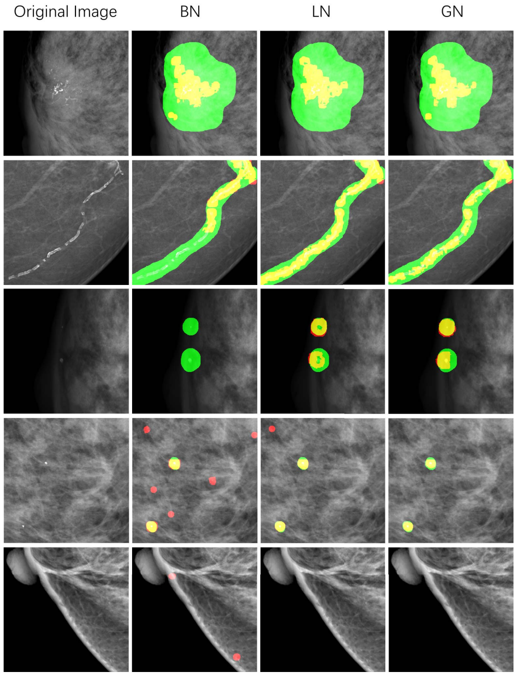


Figure 9: Comparison of calcification segmentation with different methods. From left to right, columns are original images, and segmentation results using U-Net BN, LN and GN, respectively. Green regions are the calcifications annotated by radiologists, red and yellow regions are predicted results, where yellow indicates the overlap between annotations and predictions.

5. Discussion and Future Work

We apply different deep learning frameworks for our two lesion identification tasks respectively. In the view of computer vision, a typical calcification dot is easier to be located because of its obvious contrast against the background. This implies that a simpler deep learning framework (e.g. U-Net in our approach) tends to be more robust for this task.

Authors in [9] focused on mass detection in mammograms, leveraging one public and one in-house datasets to feed the resized mammographic images along with multiple small grid patches into a one-stage RetinaNet framework. Compare to that work, the size of our public and in-house datasets are larger (Public: INbreast vs. CBIS-DDSM+INbreast+BCD; Private: 111 vs. 377 patients), and we apply the non-local operation to the two-stage detection framework Faster R-CNN.

Studies have shown that breast density has been a risk of breast cancer in women is in relation to the higher breast density. Chen *et al*. reported that when controlling for age and BMI, absolute mammographic density of Asian Americans is significantly lower than African Americans, but not compared with white women. Moreover, ethnic difference in breast density is especially significant for women older than 50 years old [2]. Investigating how much the race difference will impact the prediction performance of deep learning models is one of our next research directions.

Asymmetry and architectural distortion are two other major breast lesion types, and will be focused later when we collect more labeled data from our in-house dataset.

6. Conclusion

In this paper, we present our comprehensive solution to lesion identification in mammograms with the focus of mass detection and calcification segmentation tasks. For effectively detecting mass lesions, a Convolutional Neural Networks (CNN) based Faster R-CNN model with a series of effective modules, which are FPN, focal loss, and non-local operation, is demonstrated. Our dataset comprises three public datasets and one in-house dataset. The FROC curve shows that the Faster R-CNN model integrated with all three modules is the best model. Applying the best model validated on the public dataset, we achieve an AP of 0.933 and a recall of 0.976 as the best detection results on the in-house dataset, respectively.

To segment calcifications, we adopted U-Net with group normalization. After pre-processing is done, calcification-like artifacts were completely removed. Labels were processed to improve consistency, which benefited the model training. Moreover, a new evaluation metric was designed to address the issues raised by the imperfect labels. The experimental results have validated the efficacy of our method. In the future, we will collect more data to enhance the model and deploy it at our collaborative hospitals.

Acknowledgments

We would like to thank Dr. Le Lu and Dr. Shun Miao from PAII Inc. for their comments and suggestions. We also thank Mr. Yunhua Ji for his help in IoP implementation during his internship at PingAn Tech, US Research Lab, USA.

References

[1] Antoni Buades, Bartomeu Coll, and J-M Morel. A non-local algorithm for image denoising. In *IEEE CVPR'05*, volume 2, pages 60–65. IEEE, 2005.

[2] Zhengjia Chen, Anna H Wu, W James Gauderman, Leslie Bernstein, Huiyan Ma, Malcolm C Pike, and Giske Ursin. Does mammographic density reflect ethnic differences in breast cancer incidence rates? *American journal of epidemiology*, 159(2):140–147, 2004.

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE CVPR'09*, pages 248–255. Ieee, 2009.

[4] Jacques Ferlay, Clarisse Héry, Philippe Autier, and Rengaswamy Sankaranarayanan. Global burden of breast cancer. In *Breast cancer epidemiology*, pages 1–19. Springer, 2010.

[5] Natalie Fletcher. *Classification vs Detection vs Segmentation Models: The Differences Between Them and When to Use Each*, 2019. `https://www.clarifai.com/blog/classification-vs-detection-vs-segmentation-models-the-differences-between-them-and-how-each-impact-your-results`.

[6] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *IEEE CVPR'14*, pages 580–587, 2014.

[7] Yanan Guo, Min Dong, Zhen Yang, Xiaoli Gao, Keju Wang, Chongfan Luo, Yide Ma, and Jiuwen Zhang. A new method of detecting micro-calcification clusters in mammograms using contourlet transform and non-linking simplified pcnn. *Computer methods and programs in biomedicine*, 130:31–45, 2016.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[9] Hwejin Jung, Bumsoo Kim, Inyeop Lee, Minhwan Yoo, Junhyun Lee, Sooyoun Ham, Okhee Woo, and Jaewoo Kang. Detection of masses in mammograms using a one-stage object detector based on a deep convolutional neural network. *PloS one*, 13(9):e0203355, 2018.

[10] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[11] Rebecca Sawyer Lee, Francisco Gimenez, Assaf Hoogi, Kanae Kawai Miyake, Mia Gorovoy, and Daniel L Rubin. A curated mammography data set for use in computer-aided detection and diagnosis research. *Scientific data*, 4:170177, 2017.

[12] Rebecca Sawyer Lee, Francisco Gimenez, Assaf Hoogi, Kanae Kawai Miyake, Mia Gorovoy, and Daniel L Rubin. A curated mammography data set for use in computer-aided detection and diagnosis research. *Scientific data*, 4:170177, 2017.

[13] Peiyao Li, Zhicheng Yang, Wei Yan, Muyang Yan, Maoqing He, Qian Yuan, Ke Lan, Jiewen Zheng, Tongbo Liu, Desen Cao, and Zhengbo Zhang. Mobicardio: A clinical-grade mobile health system for cardiovascular disease management. In *2019 IEEE International Conference on Healthcare Informatics (ICHI) Workshops (AI4CDM)*. IEEE, 2019.

[14] Zhe Li, Chong Wang, Mei Han, Yuan Xue, Wei Wei, Li-Jia Li, and Li Fei-Fei. Thoracic disease identification and localization with limited supervision. In *IEEE CVPR'18*, June 2018.

[15] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *IEEE CVPR'17*, pages 2117–2125, 2017.

[16] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *IEEE ICCV'17*, pages 2980–2988, 2017.

[17] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.

[18] MA Guevara Lopez, N Gonzlez Posada, Daniel C Moura, Raúl Ramos Pollán, José M Franco Valiente, César Suárez Ortega, M Solar, Guillermo Diaz-Herrero, IMAP Ramos, J Loureiro, et al. Bcdr: a breast cancer digital repository. In *15th International conference on experimental mechanics*, 2012.

[19] Shun Miao, Z Jane Wang, and Rui Liao. A cnn regression approach for real-time 2d/3d registration. *IEEE transactions on medical imaging*, 35(5):1352–1363, 2016.

[20] Inês C Moreira, Igor Amaral, Inês Domingues, António Cardoso, Maria Joao Cardoso, and Jaime S Cardoso. Inbreast: toward a full-field digital mammographic database. *Academic radiology*, 19(2):236–248, 2012.

[21] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010.

[22] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[23] Dezső Ribli, Anna Horváth, Zsuzsa Unger, Péter Pollner, and István Csabai. Detecting and classifying lesions in mammograms with deep learning. *Scientific reports*, 8(1):4165, 2018.

[24] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[25] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *IEEE CVPR'18*, pages 7794–7803, 2018.

[26] Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–19, 2018.

[27] Fandong Zhang, Ling Luo, Xinwei Sun, Zhen Zhou, Xiuli Li, Yizhou Yu, and Yizhou Wang. Cascaded generative and discriminative learning for microcalcification detection in breast mammograms. In *IEEE CVPR'19*, pages 12578–12586, 2019.

[28] Bichen Zheng, Sang Won Yoon, and Sarah S Lam. Breast cancer diagnosis based on feature extraction using a hybrid of k-means and support vector machine algorithms. *Expert Systems with Applications*, 41(4):1476–1482, 2014.