Assessing the Impact of Deep Learning Backbones for Mass Detection in Breast Imaging

Edouard Chatzopoulos[1 [0000−0001−5848−7999]] and Sébastien Jodogne[1 (✉) [0000−0001−6685−7398]]

Institute of Information and Communication Technologies, Electronics and Applied Mathematics (ICTEAM), UCLouvain, 1348 Louvain-la-Neuve, Belgium

Abstract. Breast cancer is a significant global health concern, where early detection is crucial for effective treatment and better patient outcomes. Deep learning has emerged as a promising tool for the automated analysis of breast imaging, with the potential to reduce the workload of radiologists and enhance diagnostic sensitivity. The RetinaNet architecture is especially appealing for mass detection in mammograms, merging the advantages of single-shot detectors with the feature extraction capabilities of a feature pyramid network. This paper demonstrates the critical importance of selecting the backbone feature extractor embedded in RetinaNet. It is also shown that fine-tuning pre-trained backbones on the specific domain of interest significantly contributes to enhancing performance. This study addresses a critical gap by offering an evaluation of different feature extraction backbones and training methodologies for RetinaNet models applied to high-resolution breast imaging.

Keywords: Breast imaging · Mass detection · Deep learning · RetinaNet

1 Introduction

Breast cancer remains a significant global health concern, with early detection being a crucial factor for an effective treatment and improved patient recovery, reducing mortality by a substantial factor when implemented [33]. This process, however, is heavily reliant on the expertise of radiologists and their availability [27]. In recent years, the integration of machine learning in breast imaging has shown promising results in the automated detection of breast masses and microcalcifications. It could reduce the workload or stress experienced by radiologists due to the growing number of images they need to handle, as well as improve their sensitivity [28]. In particular, the research in the landscape of deep learning algorithms for breast mass detection has grown exponentially, reflecting the urgent need for accurate, open, and efficient diagnostic tools.

Deep learning applied to breast imaging is characterized by a wide variety of architectures, ranging from simple networks derived from natural image processing [1] to increasingly specific and complex systems [11,28,30]. However, when implementing these networks, it is common to use the default configuration of

the architecture [12,25], without considering the effect of its different components on the performance of the system.

In this paper, we evaluate the effectiveness of different variants of the RetinaNet architecture [17] for the detection of breast masses on mammograms. Indeed, the distinguishing feature of the RetinaNet architecture lies in its capability to incorporate various backbone models utilized for the binary classification of patches within images. It remains an open question as to which of these backbones yields optimal performance in breast imaging applications. Furthermore, we demonstrate the importance of adopting a learning methodology that first fine-tunes pre-trained backbones on the application domain, before training the entire RetinaNet model. These contributions offer guidelines to train simple, adaptable deep learning models whose performance reaches the state-of-the-art for mass detection in breast imaging.

2 Related Work

In this section, the main deep learning architectures that are commonly applied to mass detection in mammograms are presented. These algorithms are categorized according to the underlying architectures. Moreover, the challenges and limitations associated with the existing methods are highlighted, which will call for the novel approach that is introduced in the subsequent sections.

A first approach to computer-aided decision (CAD) based on deep learning in mammograms consists in considering mass detection as an image classification task. This restricts detection to a binary decision based on presence or absence of masses in either the entire mammogram or a predefined region of interest (ROI). This approach is simple in terms of training requirements, as it only necessitates binary annotations. However, it is far from ideal to reduce the amount of work of radiologists. Indeed, if radiologists wish to delve deeper into the decision made by the system, they have no indication about which part of the image is identified as suspicious, which necessitates further scrutiny in analysis. This lack of interpretability can be mitigated by introducing explainability components to the neural networks with techniques such as Grad-CAM [26] or weakly supervised feature localization with attention maps [7]. However, such approaches attempt to trace back the inner working of the models exclusively from the value of their existing weights and can be improved by further guiding the model with more complete annotations, including information about the localization of masses.

Another common approach consists of using segmentation-based methods, such as the well-known U-Net architecture [24]. Such methods output binary segmentation masks that closely outline the contours of the anomalies, which provides information about the localization. These methods are widely used in the context of deep learning applied to medical imaging, as they allow a precise targeting of affected tissues. In the context of breast imaging, this approach is often chosen for the detection of microcalcifications because of their wide variety in terms of shapes and dimensions, as well as because of the importance of the shape for diagnostic purposes [4]. Unfortunately, the training of segmentation-

based neural networks requires a substantially greater amount of effort compared to binary classification. Indeed, significant expertise and time are required to accurately delineate the contours of each anomaly present in the images.

An intermediate option between binary classification and segmentation-based methods is object detection. The task of object detection consists in locating bounding boxes around the elements of interest. The training of a neural network for lesion detection only requires approximate annotations provided as bounding boxes, which drastically reduces the time commitment from radiologist while maintaining a satisfactory level of explainability and localization. Work in tumor detection using bounding boxes is extensive and includes a variety of detection networks such as YOLO [21], Faster R-CNN [22], and PAA [13]. Among these methods, single-shot detection networks such as YOLO and RetinaNet are often used because of their fast inference speed. For instance, a modified version of YOLOv3 was used to obtain a model less biased towards low recall caused by the vast number of easy negative cases in mammogram databases [31]. However, YOLO networks often lack performance when it comes to the detection of small objects. Faster R-CNN, a two-shot detection network, has been proved to overcome such issues on the OMI-DB database [1]. An even deeper three-stage network based on PAA has also been shown to improve detection capabilities [11].

In this context, the RetinaNet detection network is an appealing alternative, as it combines the detection performance of Faster R-CNN with an efficient single-shot detector [17]. The RetinaNet architecture consists in incorporating a classifier network as the backbone of a feature pyramid network (FPN), which enables feature extraction at various scales. These features are fed to subsequent layers to predict the localization and class of detected targets in a single stage. RetinaNet has been used with a satisfactory level of success in literature. However, it is often restricted to images with small resolution to reduce the hardware requirements [23]. Consequently, the large resolution of mammograms (around 3000×4000 pixels) has remained a big challenge for a long time, often restraining the use of convolutional neural networks to very reduced resolutions.

To overcome this limitation, Jung et al. combined a mammogram of reduced size with 25 overlapped sections of the mammogram at full resolution [12]. The final prediction was taken as the combination of the predictions using a standard RetinaNet network with a ResNet50 backbone over the different views of the same breast. The rapid advancement of GPU capabilities, as well as the emergence of more efficient classification networks like ConvNeXt, are quickly overcoming the limitations related to the size of mammograms and can further improve the detection capabilities of RetinaNet. Unfortunately, there is currently a lack of systematic evaluation of feature extraction backbones and of training methodologies for RetinaNet models applied to breast mass detection at high resolutions. This paper aims to contribute to filling this gap. Since the release of RetinaNet, newer networks such as FreeAnchor [32] and TOOD [6] have demonstrated promising detection capabilities. However, the objective of this study is not to identify the best-performing detection network, but to explore simple architectural and training modifications that improve detection performance.

3 Materials and Methods

Designing a deep learning model for the detection of breast masses involves a multitude of decisions about the data pre-processing, about the design of the network architecture, as well as about the training methodology. This section provides a comprehensive description of the steps that were followed to train a RetinaNet model for breast imaging that integrates different backbones.

3.1 Dataset

One of the largest and most widely used mammogram datasets is the "Digital Database for Screening Mammography" (DDSM) [10]. The dataset that was used to train our RetinaNet models is the "Curated Breast Imaging Subset" (CBIS-DDSM) dataset [15], a revised and standardized subset of DDSM. This dataset is composed of 6775 digitized film mammograms from 1566 patients. The original DDSM dataset contains either imaging studies with no signs of tumor, or imaging studies with masses or microcalcifications. This contrasts with the revised version that excludes the normal exams, and only contains images of breasts with either masses or microcalcifications. Furthermore, images in the revised dataset are encoded using the DICOM interoperable format and underwent some additional corrections. In this work, only the images of the CBIS-DDSM dataset containing masses are taken into consideration. Several preprocessing steps are then applied to the source dataset, as described below:

Corrections to the dataset: Despite its high-quality data curation process, the CBIS-DDSM dataset contains some inconsistencies. Indeed, a portion of the segmentation masks delineating tumors do not match the dimensions of the mammograms. As a result, superimposing the coordinates of some masks can result in an erroneous position of the tumor on the source mammogram. In such cases, a simple resizing of the mask to the dimensions of the associated mammogram was proved to solve the issue. Consequently, in this work, wherever the source mask does not match the resolution of the mammogram, a resized version of the mask is used.

Dataset split: To prevent any contamination between the training, the validation, and the test sets, the source dataset is split at the patient level instead of the image level. This patient-wise separation of the original dataset ensures that the model is not evaluated on an image containing a mass already seen from another angle or on breast tissue from the same patient. To this end, the whole CBIS-DDSM dataset is merged, then uniformly split at the patient level according to a 80%/10%/10% train/validation/test distribution.

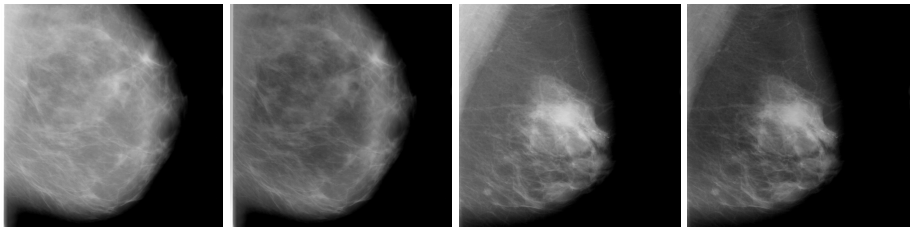
Patch generation for normal tissues: In the early phases of the training, a part of the model is trained on a patch classification task. It is therefore needed to generate 256×256 image patches that include either breast masses or normal breast tissues. Since the CBIS-DDSM dataset lacks images of normal breast tissues, such patches are extracted from all mammograms from the dataset. To avoid selecting tissue containing lesions, a segmentation mask

encompassing all anomalies in the mammogram is generated. This mask is subsequently dilated using a 256×256 pixel kernel to incorporate a safety margin around the tumors. To also avoid sampling too much background instead of healthy tissue, this dilated mask is then merged via a binary "OR" operation with a background mask, obtained through binary Otsu's thresholding of the breast image. The resulting mask is inverted to produce a segmentation mask of the normal breast tissue, with safety margins. The patch center is then uniformly drawn in this mask, to create a patch that contains normal breast tissue, while ensuring the absence of known lesions.

Patch generation for breast masses: In order to generate patches containing masses, all the masses available in the training set are randomly sampled. Two different strategies are then applied depending on the size of the sampled masses. The first strategy is used for masses with an area that is greater than 40% of the area of the patch. In this case, the center of the patch is drawn uniformly among the points of the patch situated inside the tumor. The second strategy is used for smaller masses, for which centering the patch inside the mass would result in highly similar patches when sampling the same tumor. A random perturbation is generated uniformly in the range from -25% to $+25\%$ of the patch side, in both the x and y directions. The center of the patch is then defined as the sum of the barycenter of the tumor mask and this random offset. The percentage of the area of the tumor visible on this patch is then computed and patching is repeated from the start if the ratio is smaller than 80% until this criterion is met. The combination of those two strategies ensures to combine coverage of the entirety of large masses, while benefiting from greater peripheral coverage next to smaller masses.

Data augmentation: To augment the number of training images and reduce overfitting, a combination of different data augmentation techniques is used. The images are randomly flipped with a probability of 50% in both the horizontal and vertical direction. The patches are resized with a scale uniformly distributed between 80% and 120%. Patches are rotated by an angle that is randomly selected between 0°, 90°, 180°, and 270°. Affine transformations are then applied to the patches to add an additional rotation in the $-5°$ to $+5°$ range, a relative translation in the 0% to 5% range in both directions, and a shearing in the $-5°$ to $+5°$ range in both directions. If the modified image is greater than the original dimensions, a center crop is applied with the original image dimension to reduce memory footprint of large images.

Contrast enhancement: To enhance the local contrast in dense breast tissues, each image is processed with an adaptive histogram equalization algorithm for both training and inference. The algorithm used for this purpose is Contrast Limited Adaptive Histogram Equalization (CLAHE) [20]. CLAHE is applied with a grid size of 8×8 and a clip limit of 1. To apply the algorithm, the image must first be converted to integer, and converted back in the 0 to 1 range in floating point after the contrast enhancement before being fed to the network. The effect of contrast enhancement is illustrated in Figure 1.



(a) (b) (c) (d)

Fig. 1. Examples of portions of mammograms from CBIS-DDSM before (a and c), and after (b and d) applying CLAHE.

3.2 Deep learning models

The model used in this work is an adapted version of RetinaNet [17] using different feature extraction backbones. RetinaNet is a single-stage detector network based on a feature extraction backbone extended with two different heads for classification and regression. The features of the classification network are extracted at three different scales by the feature pyramid network (FPN) [16] and fed to the two heads of the network. This multi-scale feature extraction allows combining the higher-resolution features of the first layers, which is required for the precise detection of small tumors and precise contours, while using the more abstract features of the deeper layers to obtain better classification of the lesions. The classification head of the network is responsible for detecting candidate lesions, and eventually predicting their class, while the regression head is tasked with predicting the dimensions and position of the bounding boxes.

As RetinaNet relies on its backbone to produce features relevant for lesion detection, the backbone model can have an important influence on the quality of the detection, as well as on the memory footprint of the model and on the execution time that is needed for inference. In this work, backbone with different sizes and characteristics are integrated to the network and compared on the task of detection of breast masses. The following backbones have been investigated:

ResNet: ResNet is a well-known type of classification network that is widely used on image classification tasks [9]. Its architecture is simple, using four blocks of convolutional layers and a classification head. The RetinaNet architecture is traditionally built on the top of a ResNet50 backbone.

ResNeXt: ResNeXt networks are improved versions of ResNet that exploit a new dimension, the so-called *cardinality* [29]. Compared to ResNet blocks, the blocks are reduced in the channel dimension, but each block is duplicated C times (C being the value of the cardinality) and the output of every sub-block is aggregated. Even under the condition of maintaining complexity, increasing the cardinality can improve classification accuracy of the network.

ConvNeXt: ConvNeXt networks represent another evolution of ResNet that aims to reach the same high classification performance as vision transformers [5], while maintaining the simplicity of traditional convolutional net-

Table 1. The set of hyperparameters used for the different training phases. Phases 1A and 1B consist in training the patch classifier alone. Phases 2A and 2B correspond to the training of the RetinaNet network.

Training phase	**1A**	**1B**	**2A**	**2B**
Task	Patch binary classification	Patch binary classification	Mass detection	Mass detection
Loss function	Binary cross entropy loss	Binary cross entropy loss	Focal loss	Focal loss
Trained layers	Final fully-connected layer	Complete backbone	Heads of RetinaNet	Complete RetinaNet
Batch size	32	32	1	1
Learning rate	1e-4	1e-4	1e-4	1e-5
Weight decay	5e-4	5e-4	1e-4	1e-4
Optimizer	RAdam	RAdam	SGD	SGD
Learning rate scheduler	Cosine annealing	Cosine annealing	Cosine annealing	Cosine annealing
Epochs	100	100	100	100
Batch normalization state	Unfrozen	Unfrozen	Frozen	Frozen

works [18]. Their biggest addition is a special type of convolution named "depth-wise convolution" which increases the receptive field of the network without increasing computational complexity. Because of an alternating between two distinct types of blocks inside the ConvNeXt network, connecting the FPN of RetinaNet to a ConvNeXt network is not as straightforward as with ResNet or ResNeXt networks. Indeed, instead of feeding the output of the last three blocks of the network to the FPN, we need to feed one out of every two blocks. This is because only blocks 0, 2, 4, and 6 down-sample the dimension of the feature maps, and therefore extracting two consecutive layers does not result in extracting different resolution feature maps for the different levels of the FPN. For this reason, in this work, the layers extracted by the FPN are the outputs of layers 3, 5, and 7.

3.3 Training

The training of the RetinaNet model is divided into four parts, two for training the backbone on a patch classification task, and two for training the complete RetinaNet network. The training starts with a backbone that is pre-trained on the ImageNet dataset [3]. In the first training phase (1A), all the layers of the backbone network are frozen, except for its classification head that is fine-tuned to distinguish patches containing normal tissues from patches corresponding to tumor masses. In the second phase (1B), all other layers of the backbone are unfrozen, and the model is fine-tuned on the patch classification task with a lower learning rate. Once the backbone is trained, it is extended with a FPN and integrated as the feature extractor of the RetinaNet model. In the third learning phase (2A), the resulting RetinaNet model is trained on the task of mass detection, while keeping the weights of the backbone frozen. In the final phase (2B), the weights of the backbone are unfrozen, and the learning rate is lowered again to fine-tune the entire RetinaNet network. For the last two phases, the batch normalization layers are kept in a frozen state to prevent instabilities that could result from training batch normalization with a unit batch size. Table 1 reports the hyperparameters that were used throughout the distinct phases. The four training phases are illustrated in Figure 2.

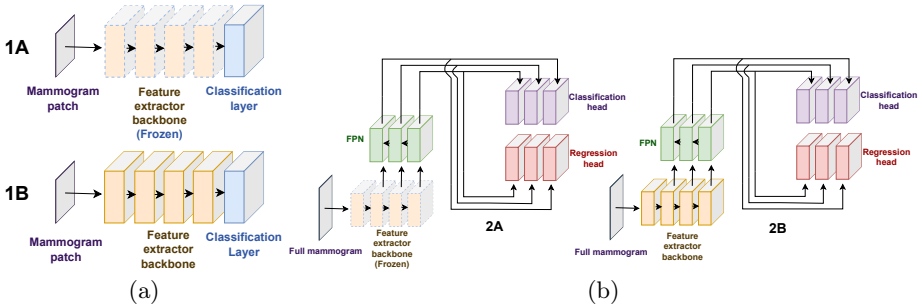


Fig. 2. Configuration of the network (a) in phases 1A and 1B, and (b) in phases 2A and 2B. The frozen layers are displayed with dotted contours and a reduced opacity.

3.4 Metrics

This study employs commonly used metrics to assess the accuracy and reliability of the networks. In the context of breast mass detection, a True Positive (TP) happens when the model correctly predicts the presence of a mass. An Intersection over Union (IoU) threshold at 0.2 between the true bounding box of a mass and the predicted one is used to consider them matching, which is consistent with standard conventions for mass detection in mammograms [11]. A False Positive (FP) happens when the model incorrectly predicts the presence of a mass. The False Positive per Image rate (FPI) is the average number of FP for a prediction on any given image. The True Positive Rate (TPR) represents the ratio of the total number of TP, against the sum of all TPs and FPs. This ratio is also known as sensitivity. Finally, the Free response Receiver Operating Curve (FROC), is the plot of sensitivity versus the average number of false positives per image. This metric is analogous to the Receiver Operating Curve (ROC) in image-level evaluation, but for lesion-level evaluation.

4 Results

In the following section, we present the results of our study on the impact of numerous factors on the performance of the breast mass detection network. We examine the effect of backbone pre-training, the influence of backbone fine-tuning within the detection network, and compare different backbone architectures for the RetinaNet model. Additionally, we analyze the performance relative to network size and compare our results with other works on the same dataset. This systematic evaluation provides insights into the optimal configurations and strategies for enhancing the accuracy of breast mass detection.

4.1 Effect of backbone pre-training

This section explores the benefits of pre-training the model used in the backbone of the detection network on the task of patch classification, instead of using a

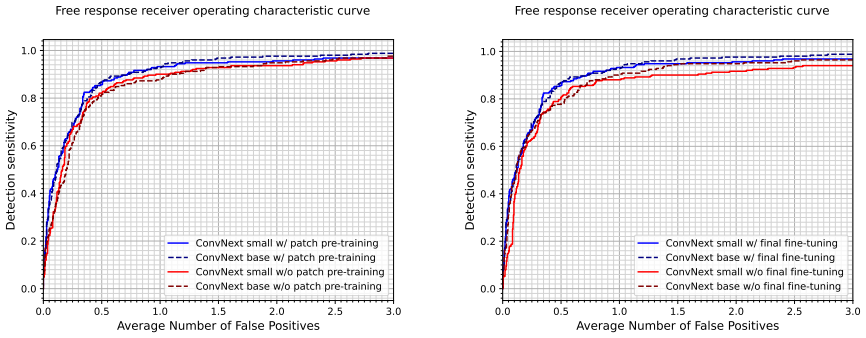


Fig. 3. (a) Effect of pre-training the backbone on a patch classification task before its integration in RetinaNet. The baseline shown in red is the result of using a backbone pre-trained on ImageNet only. (b) Effect of the addition of the final fine-tuning stage of the whole network (phase 2B) on ConvNeXt base and on ConvNeXt small. The baseline shown in red is the result of bypassing the final fine-tuning.

network that is pre-trained on ImageNet as such. Indeed, most earlier works simply fine-tune the complete RetinaNet network, without training the underlying feature extractor on the domain of interest first. Figure 3 (a) depicts the FROC curve for the RetinaNet model trained in two ways: firstly, using the four-stage approach suggested in this paper, and secondly, using only stages 2A and AB with the pre-trained backbone. A significant jump in performance can be observed at every point of the FROC curve if pre-training the backbone on the patch classification task. These results indicate that first training the backbone sub-network on the specific task of mammography mass classification can help the network converge to a better optimum in the lesion detection task. Note that both networks were trained until convergence, therefore the improvement in performance is not due to a longer training time.

4.2 Effect of backbone fine-tuning inside the detection network

The previous section indicates the interest of fine-tuning the patch classifier. One might intuitively believe that since this backbone is already fine-tuned, unfreezing its weights during the final stage of RetinaNet training would yield no additional benefits. Figure 3 (b) displays the comparison of the FROC curve of two models, before and after stage 2B. This experiment shows that even after pre-training the backbone classifier in stages 1A and 1B, the addition of stage 2B allows to further ameliorate the TPR of the network at any FPI.

This result indicates that unfreezing the weights of the backbone in the final training phase of the model is beneficial, even in the case of a feature extractor trained to convergence on the same dataset. This may be due to the additional layers corresponding to the FPN and to the two RetinaNet heads, which add

several layers of abstraction to the deeper feature of the network, hereby requiring a finer adjustment of the shallower features of the network. These findings, combined with those of Section 4.1, indicate that the optimal convergence of the model cannot rely solely on a feature extraction backbone trained on the target dataset, nor can it rely solely on the RetinaNet model to fine-tune this backbone on its own. For optimal detection performance, a combination of these two factors seems to be required.

4.3 Backbone comparison

The detection performance of the RetinaNet model depends upon the choice of the backbone. In this section, to assess the significance of this decision, various feature extraction backbones are compared for the task of detecting breast masses. Figure 4 compares the performance of RetinaNet models embedding six different backbones (ResNet50, ResNet101, RestNeXt50, ConvNeXt base, ConvNeXt small, and ConNeXt tiny) after training them using the four-stage method described in Section 3.3. This figure indicates that the traditional RetinaNet model based on a ResNet50 backbone is competitive with respect to other combinations, but its performance is overtaken by a handful of alternatives.

Among these alternatives, the models based on ConvNeXt seem to perform best, with performance consistently above the ResNet50 baseline, with a sensitivity up to 18% greater for the same FP rate. For the same FPI of 0.5, the ConvNeXt base network reaches a TPR of 0.87, while the ResNet50-based network peaks at 0.74 TPR, with an increase of 17.6% in sensitivity. ResNet101, a larger variant of the model, provides only a slightly higher overall sensitivity. However, some models that are more recent than ResNet, such as ResNeXt50, appear to be behind in terms of performance on this specific task, as can be seen by the overall lower FROC score compared to the networks configured with the other backbones. Despite the apparent advantage of the ResNeXt model in some image classification tasks [29], this superiority does not seem to transfer to mass detection in the CBIS-DDSM dataset within a RetinaNet detection model.

These findings are further validated by the mean Average Precision (mAP) of the various models at an IoU threshold of 0.2. The ResNet50-based model achieves a mAP of 0.6806, while the ResNet101-based model attains a mAP of 0.6897, indicating a marginal performance difference between the two models. In contrast, the ConvNeXt-based models exhibit significantly higher mAP values, with the tiny, small, and base variants respectively achieving mAPs of 0.7264, 0.7649, and 0.7736, which highlights their superior detection performance.

4.4 Performance relative to the size of the network

In Figure 5, a modest correlation between model size and detection capabilities can be observed, although it is not entirely consistent. Indeed, bigger models in general achieve higher scores, with the largest model demonstrating the best performance and the smallest model showing the poorest. However, the networks based on ConvNeXt small and ConvNeXt tiny exceed the performance

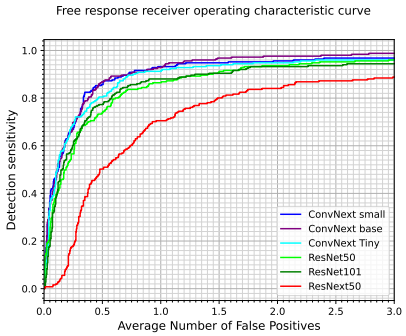


Fig. 4. Comparison of the FROC of the RetinaNet model with five different backbones. All networks are trained using the four-stage training.

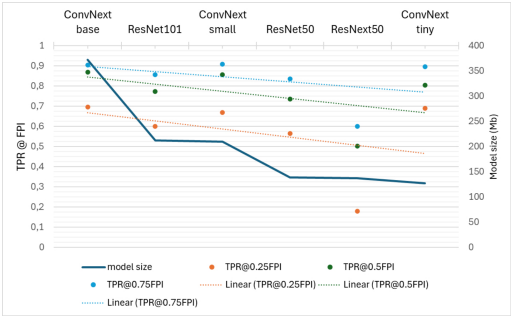


Fig. 5. Comparison of TPR for different rates of FPI in relation to total model size.

of the larger ResNet101 network at every measured threshold. This could be a result of their larger receptive field and their novel convolution method, allowing them to make better use of the large resolution of the mammogram. Moreover, the performance of RetinaNet-based on ConvNeXt small and ConvNeXt tiny is comparable to that of the ConvNeXt base model, despite the latter being larger by a factor of 77% and 192% respectively. The lack of a larger improvement in performance relatively to the size of the model indicates that with the right architecture, a small model is sufficient for mass detection on small mammogram datasets such as CBIS-DDSM. Further research would be necessary to determine whether this fact holds true for larger datasets or more complex tasks.

As expected, the inference time of the models is strongly correlated with the size of the network, with the ConvNeXt-based model respectively averaging 88ms, 143ms and 211ms per iteration for the base, small, and tiny variations on a NVIDIA RTX A4500 GPU. The ResNet-based models average 75ms and 102ms per iteration for the 50 and 101 variations on the same hardware. These inference times are all of the same order, but could be of importance in the case of inference on slower hardware.

Table 2. Comparison of the proposed method vs. CAD systems trained and evaluated on CBIS-DDSM, expressed as true positive rate at given false positive per image rate.

Method	Model type	Size of test set	TPR @ FPI
Yu et al. [30]	GFNet	348	0.89 @ 2.21 0.90 @ 2.91
Lbachir et al. [14]	HRAK + BTC + SVM	152	0.91 @ 0.65
Peng et al. [19]	Faster R-CNN + ResNeXt	Unknown	0.94 @ 2.2805
Cao et al. [2]	Faster R-CNN+FPN+FL+NL trained w/ CBIS-DDSM+INbreast+BCD	∼320	∼0.8 @ 0.4
Ours	RetinaNet + FPN + ConvNeXt base	249	0.8 @ 0.38 0.89 @ 0.57 0.95 @ 1.07

4.5 Comparison with other works on CBIS-DDSM

Table 2 shows a comparison of the proposed method against other similar CAD systems trained on CBIS-DDSM. Our method is on par with the state of the art, offering either higher TPR at comparable FPI, or similar TPR at a lower FPI. These results show that despite its size, CBIS-DDSM remains a challenging dataset. This difficulty can partly be attributed to the dataset being composed of digitized scans. These scans, originating from analog sources, exhibit lower contrast and more artifacts compared to fully digital mammograms. Access to high-quality datasets continues to be one of the major obstacles to perfecting CAD for breast cancer screening [8].

5 Conclusion

This paper introduces an efficient, comprehensive methodology to train a RetinaNet network on the task of mass detection in breast imaging. The achieved performance aligns with the state-of-the-art results on the CBIS-DDSM dataset. The importance of selecting a proper backbone network for use as a feature extractor has been highlighted. It has also been shown that fine-tuning pre-trained backbones on the specific domain of interest is a crucial element for enhancing the performance of RetinaNet models. We argue that transparently disclosing such training methodology is crucial for ensuring the reproducibility of deep learning algorithms in medical imaging.

The scope of this work is currently limited to the detection of breast masses and ignores microcalcifications or other types of lesions. Future work will train RetinaNet models that are at the same time applicable to the detection of breast masses and microcalcifications, and that predicts the malignancy of the lesions. The potential of alternative multi-scale feature extractors such as Bidirectional FPN will also be explored, as well as multi-view feature extraction to enhance the detection of challenging lesions such as architectural distortions and focal asymmetries. Finally, it is planned to publicly release pre-trained RetinaNet models as free and open-source software. This initiative aims to contribute to the advancement and implementation of computer-aided detection in breast imaging, offering benefits in education and research contexts.

References

1. 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 (2020)
2. Cao, Z., Yang, Z., Zhang, Y., Lin, R.S., Wu, S., Huang, L., Han, M., Ma, J.: Deep learning based mass detection in mammograms. In: 2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP). pp. 1–5 (2019)
3. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: IEEE CVPR 2009. pp. 248–255. IEEE (Jun 2009)
4. Dengler, J., Behrens, S., Desaga, J.: Segmentation of microcalcifications in mammograms. IEEE Transactions on Medical Imaging **12**(4), 634–642 (1993)
5. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021)
6. Feng, C., Zhong, Y., Gao, Y., Scott, M.R., Huang, W.: TOOD: Task-aligned one-stage object detection. In: ICCV 2021. pp. 3490–3499 (2021)
7. Gupta, S., Lakhotia, S., Rawat, A., Tallamraju, R.: ViTOL: Vision transformer for weakly supervised object localization. In: Proceedings of the IEEE/CVF CVPR Workshops. pp. 4101–4110 (June 2022)
8. Hassan, N., Hamad, S., Mahar, K.: Mammogram breast cancer CAD systems for mass detection and classification: A review. Multimedia Tools and Appl. **81**(14), 20043–20075 (Jun 2022)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE CVPR 2016. pp. 770–778. IEEE (Jun 2016)
10. Heath, M., Bowyer, K., Kopans, D., Moore, R., Kegelmeyer, W.P.: The digital database for screening mammography. In: Proceedings of the 5th International Workshop on Digital Mammography. pp. 212–218 (2001), `http://www.eng.usf.edu/cvprg/Mammography/Database.html`
11. Jiang, J., Peng, J., Hu, C., Jian, W., Wang, X., Liu, W.: Breast cancer detection and classification in mammogram using a three-stage deep learning framework based on paa algorithm. Artificial Intelligence in Medicine **134**, 102419 (2022)
12. Jung, H., Kim, B., Lee, I., Yoo, M., Lee, J., Ham, S., Woo, O., Kang, J.: Detection of masses in mammograms using a one-stage object detector based on a deep convolutional neural network. PLOS ONE **13**(9), 1–16 (09 2018)
13. Kim, K., Lee, H.S.: Probabilistic anchor assignment with iou prediction for object detection. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (eds.) ECCV 2020. pp. 355–371. Springer International Publishing, Cham (2020)
14. Lbachir, I.A., Daoudi, I., Tallal, S.: Automatic computer-aided diagnosis system for mass detection and classification in mammography. Multimedia Tools and Appl. **80**(6), 9493–9525 (Mar 2021)
15. Lee, R., Gimenez, F., Hoogi, A., Miyake, K., Gorovoy, M., Rubin, D.: A curated mammography data set for use in computer-aided detection and diagnosis research. Scientific Data **4**(1) (Dec 2017)
16. Lin, T.Y., Dollar, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: IEEE CVPR 2017. pp. 936–944. IEEE (Jul 2017)
17. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence **42**(2), 318–327 (2020)

18. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A ConvNet for the 2020s. In: IEEE/CVF CVPR 2022. pp. 11966–11976 (2022)
19. Peng, J., Bao, C., Hu, C., Wang, X., Jian, W., Liu, W.: Automated mammographic mass detection using deformable convolution and multiscale features. Medical & Biological Engineering & Computing **58**(7) (Jul 2020)
20. Pizer, S.M., Amburn, E.P., Austin, J.D., Cromartie, R., Geselowitz, A., Greer, T., ter Haar Romeny, B., Zimmerman, J.B., Zuiderveld, K.: Adaptive histogram equalization and its variations. Computer Vision, Graphics, and Image Processing **39**(3), 355–368 (1987)
21. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: 2016 IEEE CVPR. pp. 779–788 (2016)
22. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. In: Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 28. Curran Associates, Inc. (2015)
23. Ribeiro, R.F., Torres, H.R., Oliveira, B., Morais, P., Vilaça, J.L.: Comparative analysis of deep learning methods for lesion detection on full screening mammography. In: 2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). pp. 1–4 (2023)
24. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. pp. 234–241. Springer International Publishing, Cham (2015)
25. Ryspayeva, M., Molinara, M., Bria, A., Marrocco, C., Tortorella, F.: Transfer learning in breast mass detection on the OMI-DB dataset: A preliminary study. In: Rousseau, J.J., Kapralos, B. (eds.) Pattern Recognition, Computer Vision, and Image Processing. ICPR 2022 International Workshops and Challenges. pp. 529–538. Springer Nature Switzerland, Cham (2023)
26. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: IEEE ICCV 2017. pp. 618–626 (2017)
27. Sickles, E.A., Wolverton, D.E., Dee, K.E.: Performance parameters for screening and diagnostic mammography: Specialist and general radiologists. Radiology **224**(3), 861–869 (2002), pMID: 12202726
28. Wu, N., Phang, J., Park, J., Shen, Y., Huang, Z., other: Deep neural networks improve radiologists' performance in breast cancer screening. IEEE Transactions on Medical Imaging **39**(4), 1184–1194 (2020)
29. Xie, S., Girshick, R., Dollar, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: IEEE CVPR 2017. pp. 5987–5995. IEEE Computer Society, Los Alamitos, CA, USA (jul 2017)
30. Yu, X., Zhu, Z., Alon, Y., Guttery, D.S., Zhang, Y.: GFNet: A deep learning framework for breast mass detection. Electronics **12**(7) (2023)
31. Zhang, L., Li, Y., Chen, H., Wu, W., Chen, K., Wang, S.: Anchor-free YOLOv3 for mass detection in mammogram. Expert Systems with Appl. **191**, 116273 (2022)
32. Zhang, X., Wan, F., Liu, C., Ji, R., Ye, Q.: FreeAnchor: Learning to match anchors for visual object detection. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) NeurIPS 2019. vol. 32. Curran Associates, Inc. (2019)
33. Zielonke, N., Gini, A., Jansen, E.E., Anttila, A., Segnan, N., et al.: Evidence for reducing cancer-specific mortality due to screening for breast cancer in europe: A systematic review. European Journal of Cancer **127**, 191–206 (2020)