A mammography classification model trained from image labels only

Fredrik A. Dahl*[1], Marit Holden[1], Olav Brautaset[1], and Line Eikvil[1]

[1]Norwegian Computing Center, Oslo, Norway

Abstract

The Cancer Registry of Norway organises a population-based breast cancer screening program. The interpretation of the screening mammograms is a manual process, but deep neural networks are showing potential in mammographic screening. Most methods focus on methods trained from pixel-level annotations, but these require expertise and are time-consuming to produce. Through the screenings, image level annotations are however readily available. In this work we present a few models trained from image level annotations from the Norwegian dataset: a holistic model, an attention model and an ensemble model. The models combine two ResNet101 instances and use gradient based saliency mappings to identify regions of interest. We compared the models' performance with that of pretrained models based on pixel-level annotations, trained on international datasets. From this we found that models trained on our local data with image-level annotation gave considerably better performance (AUC=0.940) than the pretrained models from external data (AUC=0.861), although the latter was based on pixel-level annotations.

1 Introduction

Breast cancer is the most common cancer and the leading cause of cancer death among women worldwide [3]. Early detection of breast cancer through screening is recommended by international health organizations to reduce this mortality [11]. In Norway a population-based breast cancer screening program [8], administered by the Cancer Registry of Norway (CR) invites all women aged 50-69 to mammography every two years. About 250 000 women participate each year.

The interpretation of the screening mammograms is a manual process where two radiologists independently analyze the mammograms, giving them a score indicating the level of suspicion of malignancy. If both radiologists give the lowest score, the screening is considered negative, otherwise a consensus meeting is used to determine whether to call the woman back for further assessment. In total 0.6% of the attending women are diagnosed with breast cancer.

Artificial intelligence (AI) building on deep neural networks is now showing potential in mammographic screening and may help reduce the workburden of radiologists. Initial AI approaches [2, 6] were hindered by limited amount of training data, but more recent studies [20, 13] have had much more data available, and some report performance that competes with radiologists [13].

Compared to the more traditional methods for mammogram analysis, deep learning methods offer the advantages of automated learning of features and the opportunity for end-to-end models. However, to achieve this, a considerable amount of annotated training data is needed. Through the Norwegian breast cancer program, we have a large dataset available. However, we only have diagnosis and annotations at image and patient level, and not delineation of cancers at pixel level. Acquiring this type of pixel level annotations is time consuming and requires breast radiologists that are already overloaded with work. Our choice was therefore either to train from our own dataset based on image-level labels or to use an existing state-of-the-art model that was pre-trained on pixel information directly or for transfer learning.

*Corresponding Author: fadahl@nr.no

https://doi.org/10.7557/18.6244

2 Related work

As mammographic images are large, full resolution images will typically not fit into GPU memory together with the more accurate standard CNN architectures. At the same time, the salient features of a growing tumor, such as lesions and calcifications, can only be reliably identified on a very small scale. Therefore, significant information could be lost from down-sampling, resulting in poor performance [15]. Most approaches are therefore based on methods that need to be trained from pixel-level annotations [9, 10, 4].

Still there are some works on holistic approaches, using image-level annotations only [14, 22, 6]. Mohammad et al [14] obtain a holistic approach through inputting a heavy down-sampled mammogram to a CNN. Zhu et al [22] perform classification based on a raw, un-annotated whole mammogram, where they attempt to overcome the resolution problem through using a multiple instance classifier approach based on a bag-of-patches. Geras et al [6] succeed at using high-resolution input in their multi-view deep convolutional network (MV-DCN), through using a specially designed architecture with aggressive convolution and pooling layers. Through this they are able to build an MV-DCN that takes four 2600x2000 pixels images (one per view) as input without any downscaling and without much performance degradation.

Wu et al [20] investigate the value of pixel-level labels by comparing the use of only image-level labels to that of using both image-level and pixel level labels, where they find that the combination performs better than the image-only model. Their image-only model is inspired by the method of Geras [6]. Pixel-level and image-level labels are combined by using a pixel-trained model to predict heatmaps which are then used to produce multi-channel images as input to the image-level model. Still, they comment that the high resolution of the images and the limited memory of modern GPUs constrain them to use relatively shallow networks (ResNet-22) within the models when using full-resolution images as inputs.

To be able to exploit information from local details without having pixel-level annotations, more recent works propose to use weak supervision [15, 16, 12]. Such approaches aim to identify image regions relevant to classification utilizing only image-level labels during training, based upon the observation that feature maps in the final convolutional layers of CNNs reveal the most influential regions of the input image [12].

Shen et al have developed what they call a globally aware multiple instance classifier (GMIC) [15]. They use the ResNet-model of Wu et al to generate a pixel-level saliency map through a 1x1 convolution of the feature maps obtained after the last residual block. Each element of the saliency map denotes a score that indicates the contribution of the corresponding location towards classifying the input. The most salient patches are then selected through a Gated Attention Mechanism and sent to a multiple instance classifier. In [16] they extend the architecture with a fusion model. This approach requires image-level annotations only and compares favourably to the approaches of Wu in terms of performance but requires more memory and computational resources. As a biproduct of the classification procedure, these models can visualize the important patches.

Liu et al [12] have introduced another weakly supervised approach that also aims at segmentation, GLAM (Global-Local Activation Maps). In comparison with GMIC, this approach produces more fine-grained saliency maps. They also combine saliency maps at different scales to generate the global saliency map. With their model they also provide segmentation. Classification-wise their AUC is on level with the GMIC model.

In our work we wanted to first look into how far a simple holistic approach could bring us, when we were able to do end-to-end training on our Norwegian dataset and while finding a good balance between input resolution and network capacity. We compared this to a pre-trained model trained with a combination of global and local information. Furthermore, to see if our holistic model could gain from incorporating even more detailed information, we also extended our baseline model into a simple attention model. In our attention model, we focused on capturing salient areas at high resolution, while keeping patches as large as possible and reusing the network architecture from the holistic model.

3 Methods

In the following we describe the methods based on training from image-level annotations, and also give a brief description of the method behind the pre-trained model that also exploited pixel-level annotations.

3.1 Holistic model

In our first approach we wanted to see how a simple holistic approach would fare, that exploits only image-level annotations and could be trained end-to-end on our dataset. We chose a model of the residual net family, which have been much used for mammogram analyses [20, 12, 15, 16]. They show good performance and are neither too memory nor too computation demanding. We found that a ResNet-101 network gave us a good balance between network capacity and memory needs, allowing for use of image crops of good size and reasonable resolution although some downsampling was still required.

For the purpose of model training, we defined the positive cases as examinations for which cancer was found through the screening procedure, and the rest of the examinations as negative ones. Input to the network was a single image, and standard data augmentation transforms were used during training. In evaluation mode, the images were center-cropped and down-sampled before they were fed into the network. The network output was a positive and a negative class score, where we defined a risk score for cancer for a single input image as the difference between these (positive - negative).

3.2 Attention model

Even without pixel-level annotations we still wanted to try to identify important regions in the image and see if our ResNet model could gain from increased attention towards suspicious regions. To achieve this we used the simple pixel attribution approach of Simonyan et al [17], where they calculate the gradient of the loss function for the class they are interested in with respect to the input pixels. This model has been shown to give reasonable results on medical images [7] and it is computationally efficient, since it requires only a single back-propagation pass through the model.

By retrieval of the gradients from predictions from our original holistic model, we could then get a spatial attention map showing where the network focused in order to classify the given image. From this attention map we then identified the most salient parts of the image by finding the points with the highest gradient impact. From these points we found the mass centre and defined this as the centre of attention (COA) in the image and identified a region of interest (ROI) around this COA. Figure 1 shows an example of saliency maps and resulting ROIs.

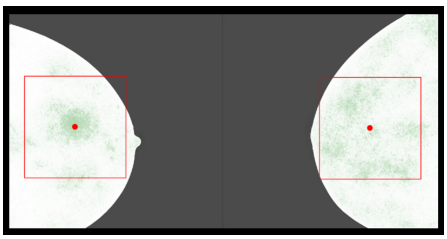

Figure 1: Examples of saliency maps for a positive (left) and negative (right) case. The green dots represent the important pixels, the red dot shows the COA and the squares are the ROIs.

Under the assumption that the important part of the image has been found, the region of interest can be considerably smaller than the full image and can thereby be represented with a higher resolution and still fit on the GPU. One might suspect that this method would fail to identify relevant parts of positive images that were falsely classified as negative by the Holistic model, but subsequent performance tests indicate that it performs reasonably well.

Since the Holistic model was designed to utilize the available GPU resources as effectively and efficiently as possible, we decided to re-use the same architecture for the fine-grained model; the only difference being that the fine-grained model was trained to evaluate a smaller part of the image and therefore received input of a higher resolution, with the aim of improving its ability to identify minute indications of cancer.

This attention approach has some similarities with the approaches in [16, 12], by requiring image-level labels only for training and using a single holistic ResNet for identifying ROIs to be passed on to a more fine-grained sub-model. The main difference lies in the degree of complexity. While

the GLAM model [12] uses a hierarchy of patches of different resolutions and the GMIC model [16] uses non-standard loss functions, tailor-made optimization procedures for selection of patches and attention mechanisms that combine model outputs from several ROI patches, our approach is simpler and more transparent.

3.3 Ensemble model

While we would expect the focused attention and higher resolution of the Attention model to lead to a better performance than that of the simpler Holistic model, this might not always be the case. For some images, the over-all shape of the breast or other macro-characteristics might give independent information about the correct diagnosis, which would be lost for the fine-grained sub-model. There may also be several local regions of a breast that contribute independently to a positive evaluation by the Holistic model, in a way that could be lost in the ROI. Besides, there is often a benefit from averaging the output of different models [5], and for these reasons we defined an Ensemble model through the (unweighted) average of the Holistic and Attention risk scores.

3.4 Examination-level risk scores

A screening examination consists of four images with two different views for each breast. To obtain the risk score of a screening examination, we first defined a single-breast risk score as the sum of the risk scores for the two images of the given breast, and then we defined the examination-level risk score as the maximum of the two single-breast scores. This procedure was applied to the single-image risk scores of the Holistic, Attention and Ensemble models alike.

3.5 Pretrained model

For our Pretrained model we wanted one that was trained from a combination of image-level labels and pixel-level annotations and could be adapted to our image-level labels. The model described in [20] was chosen, as it fitted this order and also had implementation code and pretrained weights available on GitHub. This model uses a multi-view approach that takes as input all four views from a screening examination. To obtain local information, heatmaps are first estimated for each view using a DenseNet model trained on smaller annotated image patches. For each view, this is used to produce heatmaps for likely benign and malignant masses. A multi-view model takes these heatmaps as input, together with the actual images, and produces two risk scores (benign and malignant) for each breast. For further details we refer to [20]. We used the malignancy risk scores of the Pretrained model directly and also implemented transfer learning. For the latter, we replaced the head of the pretrained model with a head adapted to and trained on our own Norwegian data set, predicting cancer at examination-level.

4 Experiments

4.1 Data

Our data set was from mammography screening of women aged 50 to 69 in breast-diagnostic centres from three Norwegian counties, containing data from a total of 153 159 examinations using Siemens equipment.

Each examination in our data set consisted of four grey-scale images; two of each breast taken from two different angles (Figure 2). The four images are from left to right L-CC, L-MLO, R-CC and R-MLO, where R and L means right and left breast, respectively, while MLO (mediolateral oblique) and CC (bilateral craniocaudal) are the two standard views of the breast. Some examinations had more than four images, but these were excluded from the analysis. There were eight different image sizes ranging from 940x2340 to 3328x4096 pixels.

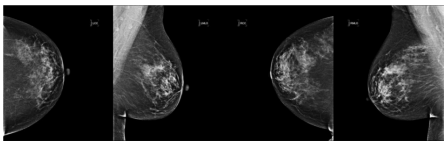

Figure 2: An example of four-view mammographic images (not from our dataset). Illustration from The Digital Mammography DREAM challenge [18].

In our dataset we have annotations resulting from diagnoses. A cancer case is defined as a 'screening cancer' if it was diagnosed as the result

of a screening examination. Otherwise, it is defined as an 'interval cancer' because it was found and diagnosed in the time span between one screening exam and the planned next one. They are likely a mix of cancers that develop quickly after a screening exam and cases that could have been identified in screening but were missed. In some cases, they may also have been identified by the radiologists as suspicious but not confirmed by biopsies. The remaining examinations were denoted as 'normal'. In our data set of 153 159 examinations, 783 were classified as screening cancers and 249 as interval cancers.

During training, screening cancer examinations were defined as the positive cases, while the rest (including interval cancers) were defined as negative ones. For the positive cases, the two images of the affected breast were defined as positive examples, while for negative cases all four images were defined as negative ones.

4.2 Implementation details

For our Holistic model, we used a Pytorch ResNet with two output nodes, initialized with pretrained weights and modified to greyscale. We used one of the cross-validation folds to test different tradeoffs between input field size, batch size and model size within the constraints of the available GPU memory capacity (11178 MiB). We ended up with using an input field of 976x976 pixels, a batch size of 4 and a ResNet101. During training, the images were rescaled to a width of 976, padded with zeros and randomly flipped, rotated and cropped. More advanced augmentation methods like perspective transformations and elastic deformations might have been appropriate as well, but were not implemented in the present application. In evaluation mode, the images were rescaled to a width of 976 and center cropped.

The Attention model used the same crop size as the Holistic model, although of a smaller area with higher resolution. When selecting the region of interest, the centroid of the 1% pixels with the highest gradient impact was defined as the Holistic model's center of attention (COA). This percentage was defined based on visual inspection of what appeared to give reasonable results. A 976x976 square centered at the COA was used as input field for the Attention model.

The network, ResNet-101, was structurally identical to the Holistic model and the input transforms of the ROI for training were also applied with minor adjustments. Even though it evaluates on a smaller scale, initializing the Attention model with the Holistic model weights boosted the training speed substantially.

4.3 Experimental setup

All experiments were performed with 10-fold cross-validation, stratified by screening and interval cancers. As our performance metric we used the generic criterion of AUC, which is common in the field and has the advantage of being independent of class imbalance. Based on the radiologists' initial malignancy scores, we defined the following risk scores: For a single radiologist, we defined his examination-level risk score as the maximum of the single-breast scores. The double-radiologist risk score of a single breast we defined as the sum of the radiologists' risk scores, and the maximum of these was used as the examination-level risk score.

5 Results

The different models were tested separately on the same dataset. Performance was investigated both for the case of separating screening cancers from all others, and the case of separating all cancers, both screening and interval, from the rest. These results are summarized in Table 1.

From this we see that the best performance is achieved by the ensemble model, combining the Holistic and the Attention model. Furthermore, we see that all the locally trained models outperform the pretrained ones, even though the former have not been trained on pixel-level information.

In comparison with the radiologists, we also see quite good performance of the locally trained models. For the case of distinguishing between screening cancers and others, the best model, the Ensemble model, shows slightly better performance than one single radiologist. When it comes to distinguishing between all cancers (screening+interval) and others, the Ensemble model is actually on par with the double-radiologist performance.

Model	Screening	Screening + Interval
Holistic	0.913	0.866
Attention	0.932	0.883
Ensemble	**0.940**	**0.891**
Pretrained	0.833	0.784
Adapted Pretrained	0.861	0.819
1 radiologist	0.930	0.836
2 radiologists	**0.987**	**0.886**

Table 1: AUC-values for (i) screening cancer vs the rest and (ii) cancers (screening + interval) vs the rest for different models and radiologist evaluations. Note that in the AUC computations we included for efficiency reasons, all the (screening) cancers, but only 5000 of the approximateley 150 000 examinations in the rest dataset.

6 Discussion

Our main finding is that training a model end-to-end on local data with image-level labels gave better performance than a model that was pretrained on external images with pixel-level annotations. We regard this as a robust finding since even the Holistic model performed better than the version of the Pretrained model where the top layer was replaced and trained on our data. From other studies, generalization to new data sets is known to give problems [19] which may e. g. be related to different technical imaging machinery, different machine settings and physiological differences in the screened population.

We find it encouraging that even the Holistic model performs better than a single radiologist in identifying cancer cases over-all, as it is often argued that fine-grained models that classify small patches are necessary to capture the minute cancer signs present in high-resolution mammograms [21, 16]. Further improvement is seen with the Attention model, where a focus for attention is found only from the weak image-level labels. The best results are achieved with the combination of the two in the Ensemble model, combining both the holistic and the fine-grained view.

The Ensemble model has an AUC performance on screening cancers that is comparable to contemporary state-of-the-art models [1]. Such comparisons may be less meaningful across datasets, since the different models are trained and evaluated on different populations of women and images may have different technical quality. However, the finding is strengthened by the fact that the model performance was between that of a single and a double radiologist reading for detecting screening cancers, which has also been reported for other studies [1].

Furthermore, the model performance is equal to a double radiologist reading in identifying cancers in general, including screening and interval cases. This is an even more interesting finding than the detection of screening cancers. Since the radiologist scores trigger the further examinations and biopsies, the fact that an examination is given a high radiologist score increases the likelihood that it ends up in the screening cancer category. Conversely, any cancer case that might be overlooked by the radiologist is excluded from the screening cancer set. Therefore, the task of screening cancer detection by design favours the radiologists.

7 Conclusion

We conclude that the ability to fully train a model on local images gave a big advantage over an external model trained with pixel-level annotations, even when the latter was adapted using transfer learning for local images. A pure ResNet model performed surprisingly well and a model where that ResNet identifies an ROI for another model, worked even better. The latter achieved a performance comparable to double radiologist reading, for the task of identifying any cancer (screening or interval type), requiring no pixel-level annotation for training.

Acknowledgement

This research was funded by the Research Council of Norway under the projects MIM (#282040), AIforScreening (#322211) and Visual Intelligence (#309439) in cooperation with the Cancer Registry of Norway and their mammographic screening program.

References

[1] L. Abdelrahman et al. Convolutional neural networks for breast cancer detection in mammography: A survey. *Comput. Biol. Med. 113*, 2021. doi: 10.1016/j.compbiomed.2021.104248.

[2] G. Carneiro et al. Unregistered multiview mammogram analysis with pre-trained deep learning models. In *Int. Conf. on Medical Image Computing and Computer-Assisted Interv*, pages 652–660, Springer, Cham., Oct, 2015. doi: 10.1007/978-3-319-24574-4_78.

[3] J. Ferlay et al. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *International journal of cancer*, 136(5):E359–E386, 2015. doi: 10.1002/ijc.29210.

[4] T. Févry et al. Improving localization-based approaches for breast cancer screening exam classification. In *Proceedings of Machine Learning Research*, Extended Abstract MIDL 2019.

[5] M. A. Ganaie et al. Ensemble deep learning: A review, 2021. arXiv preprint arXiv:2104.02395v1.

[6] K. J. Geras et al. High-resolution breast cancer screening with multi-view deep convolutional neural networks, 2017. arXiv preprint arXiv:1703.07047.

[7] Gulum et al. A review of explainable deep learning cancer detection models in medical imaging. *Applied Sciences*, 11(10), 2021. doi: 10.3390/app11104573.

[8] S. Hofvind et al. The norwegian breast cancer screening program, 1996-2016: Celebrating 20 years of organised mammographic screening, 2017. In: Cancer in Norway 2016 - Cancer incidence, mortality, survival and prevalence in Norway. Oslo: Cancer Registry of Norway.

[9] H. E. Kim et al. Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study. *The Lancet Digital Health*, 2.3: e138–e148, 2020. doi: 10.1016/S2589-7500(20)30003-0.

[10] T. Kooi and N. Karssemeijer. Classifying symmetrical differences and temporal change for the detection of malignant masses in mammography using deep neural networks. *Journal of Medical Imaging*, 4.4:044501, 2017. doi: 10.1117/1.JMI.4.4.044501.

[11] B. Lauby-Secretan et al. Breast-cancer screening — viewpoint of the IARC working group. *N Engl J Med*, pages 372:2353–2358, 2015. doi: 10.1056/NEJMsr1504363.

[12] K. Liu et al. Weakly-supervised high-resolution segmentation of mammography images for breast cancer diagnosis, 2021. arXiv preprint arXiv:2106.07049.

[13] S. M. McKinney et al. International evaluation of an AI system for breast cancer screening. *Nature*, 577(7788):89–94, 2020. doi: 10.1038/s41586-019-1799-6.

[14] C. Mohammad et al. Breast cancer detection in mammograms using convolutional neural network. In *2018 International Conference on Computing, Mathematics and Engineering Technologies*, 2018. doi: 10.1109/ICOMET.2018.8346384.

[15] Y. Shen et al. Globally-aware multiple instance classifier for breast cancer screening. In *International Workshop on Machine Learning in Medical Imaging*, pages 18–26, Springer, Cham., 2019. doi: 10.1007/978-3-030-32692-0_3.

[16] Y. Shen et al. An interpretable classifier for high-resolution breast cancer screening images utilizing weakly supervised localization. *Medical image analysis*, 68:101908, 2021. doi: 10.1016/j.media.2020.101908.

[17] K. Simonyan et al. Deep inside convolutional networks: Visualising image classification models and saliency maps, 2014. arXiv:1312.6034.

[18] Url. `https://www.synapse.org/#!Synapse:syn4224222/wiki/401743`.

[19] X. Whang. Inconsistent performance of deep learning models on mammogram classification. *Journal of the American College of Radiology*, 17(6):796–803, 2020. doi: 10.1016/j.jacr.2020.01.006.

[20] N. Wu et al. Deep neural networks improve radiologists' performance in breast cancer screening. *IEEE transactions on medical imaging*, 39(4):1184–1194, 2019 Oct 7. doi: 10.1109/TMI.2019.2945514.

[21] C. Zhang et al. New convolutional neural network model for screening and diagnosis of mammograms. *PLOS ONE*, 15(8), 2020. doi: 10.1371/journal.pone.0237674.

[22] W. Zhu et al. Deep multi-instance networks with sparse label assignment for whole mammogram classification. In *Int. Conf. on Medical Image Computing and Computer-Assisted Intervention*, Springer, Cham, 2017. doi: 10.1007/978-3-319-66179-7_69.