NORDIC MACHINE INTELLIGENCE

Nordic Machine Intelligence, vol. 3, pp. 1-7, 2023
Received 12 Aug 2023 / published 17 Nov 2023
https://doi.org/10.5617/nmi.10459

A two-stage mammography classification model using explainable-AI for ROI detection

Fredrik A. Dahl[1*], Olav Brautaset[1], Marit Holden[1], Line Eikvil[1], Marthe Larsen[2], Solveig Hofvind[2]

1. *Norwegian Computing Center, Oslo, Norway*
2. *Section for Breast Cancer Screening, Cancer Registry of Norway, Oslo, Norway*

** E-mail to: fadahl@nr.no*

Abstract

This study introduces an enhanced version of a two-stage modelling approach using artificial intelligence (AI) for breast cancer detection in mammography screening. Leveraging a large dataset of 2,863,175 mammograms from the BreastScreen Norway, the approach uses two convolutional neural networks. The first one is trained to classify whole images, and an explainable-AI method is applied to this network to identify a region of interest (ROI). The second neural network subsequently classifies the ROI for malignancy. While a prior method used simple gradient saliency maps to identify ROIs, a key enhancement of the present methodology is the application of Layered GradCam, which identifies cancerous areas more consistently and allows smaller ROIs. Layered GradCam is also used to display identified cancers to the user. By the AUC criterion, our model performs well, 0.974 for screen-detected and 0.931 for all cancers (screen-detected and interval), compared to a commercial program; 0.959 and 0.918, respectively. Comparisons with the radiologist scores indicate that the model has equal performance with two radiologists, and superior performance to one, for the detection of all cancers (screening- and interval type). Our tests indicate that our model generalizes well for different breast centers, but so far only images from a single manufacturer have been tested.

Keywords: artificial intelligence; machine learning; mammography; explainable AI

Introduction

International health authorities advocate mammography screening to detect breast cancer early and reduce mortality from the disease (1). The Cancer Registry of Norway administers the national program, BreastScreen Norway, where about 250,000 women participate annually (2).

The screening program employs a manual process, where two radiologists independently interpret mammograms and assign a malignancy suspicion score. If both radiologists assign the lowest score, the screening result is deemed negative. Otherwise, a consensus meeting determines whether further assessment is necessary.

Deep neural networks show promise in mammography screening, and commercial systems already exist. Previous research with access to larger datasets suggests AI performance can match that of radiologists (3). This has been confirmed through an intermediate analysis of an on-going Swedish study, where AI-supported interpretation found more cancers than two radiologists (4).

Compared to traditional mammogram analysis techniques, deep learning provides automated feature learning and the opportunity for more accurate models. However, it demands a substantial amount of annotated training data. Through BreastScreen Norway, we have access to a sizeable dataset. Still, our dataset only includes diagnoses and image or patient level annotations, not pixel-level cancer delineations. Acquiring these pixel-level annotations is labor-intensive and burdens the already overloaded radiologists. We have therefore developed a methodology for training classifiers based on image-level labels, only.

The present article presents a refined version of a two-stage modelling approach that was introduced in (5). We use an improved procedure for identifying regions of interest (ROIs) and can report on substantially improved performance.

Materials and methods

Our data set was collected from 7 breast centers (BCs) across Norway over the period 2007 – 2018. Women aged 50-69 were invited to screening biennially. A total of 249,110 women are represented in the data set. We had access to pseudonymized ids, so that individual women could be traced, and the number of examinations per woman ranged from 1 to 8. The total number of examinations was 689,910.

Screen-detected cancer (n=4,105) was defined as cancer detected after a positive screening examination and a subsequent assessment. In addition, there were 1,199 cancers screened negative, but detected between an examination and the planned next one, called interval cancers. All cancer cases were histologically verified. In 131 examinations, cancer was found in both breasts, in which case the side with the primary tumor was defined as the side with cancer. There was no information on cancer location beyond left or right breast.

Table 1 gives the distribution of screen-detected cancers, interval cancers and negative examinations by age groups, while Table 2 gives the same counts by BDS.

Age at screening	Screen-detected	Interval	Negative
< 55	858	303	171,283
55-59	829	289	175,927
60-64	1098	293	170,417
> 64	1320	314	166,979

Table 1: diagnosis by age at screening.

Breast center	Screen-detected	Interval	Negative
1	574	179	105,294
2	463	146	72,602
3	322	67	52,671
4	652	186	115,831
5	328	133	63,433
6	998	276	145,724
7*	768	212	129,051

Table 2: diagnosis by BC, *=Østfold.

In the manual process, each examination was routinely interpreted by two independent radiologists, where both gave an integer risk score of 1 to 5 for each breast, where a score of 2 or higher indicated suspicious findings requiring discussion in a consensus meeting.

A standard examination includes a craniocaudal (CC) view and a mediolateral oblique (MLO) view of each breast, for a total of four images. In some cases, one or more views were missing, but more often additional images were included, e.g. retake or particularly large breasts. The total number of mammograms was 2,863,175, for an average of 4.15 per examination. The mammograms were all digital x-ray images produced with Siemens equipment. The image size varied, with 2082x2800 pixels the most common and the rest mainly 2800x3500.

All examinations carried out at a single breast center (Østfold) were held out as a test set, and not accessed during any phase of the model development. The geographical mobility of women in the given age group in Norway is low, and only 2.7% of these examinations came from women who were also screened in other centers.

The remaining data was randomized into five folds, with examinations of a given woman in the same fold. The folds were simultaneously stratified for an even distribution of screen-detected cancers, interval cancers and BDSs.

In the training procedure we applied five-fold cross-validation (6), which means that five separate models (actually *model-pairs*, as described below) were trained. Each of these were blinded to a single fold and trained on the four remaining ones. This design has the advantage that after training is completed, each mammogram can be evaluated by a model that did not see it during training. We can therefore provide sound model performance estimates on the full set of five folds. Also, the set of five slightly different models can be utilized as an ensemble for evaluating mammograms in the test set, when this is finally opened.

In the following, we first give an outline of a model published previously (5), before we describe the improvements made in the present one. A resnet101 (7) was trained to classify full-size mammograms that were center-cropped and down-sampled to 976x976 pixels. The net was initialized with weights pretrained on ImageNet. Cross-entropy loss was used with screen-detected cancer as positive and the rest as negative cases. This included interval cancers as negative training examples, which could be a problem. It is, however, a minor issue due to the large number of negative images and the balanced sampling explained below. The network output gave a positive and negative class score, and we defined a risk score as the difference between these (positive – negative).

Given the low rate of cancers in mammography screening, the dataset was extremely imbalanced, with only 0.3% positive images. The standard training approach of cycling through the data set would therefore be inefficient, so instead we implemented balanced sampling. This amounts to alternating between randomly drawn positive and negative images, with replacement.

Although our data set was large, there was considerable benefit from data augmentation. Figure 1 from (5) illustrates how this was done. The leftmost image was the original mammogram in full resolution, while next was down-sampled to a width of 976 and zero-padded. This image was randomly flipped and rotated, followed by random cropping down to 976x976.

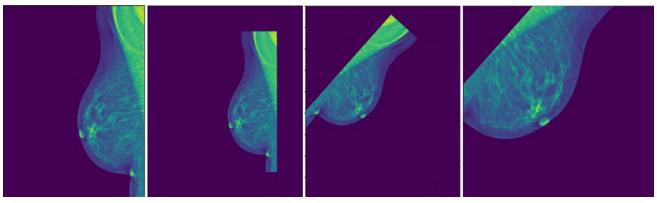

Figure 1: Data augmentation through random flipping, rotation and cropping.

While the Holistic model was able to separate positive and negative images rather well, we were able to improve on this through a two-stage design. We used a gradient-based method that traced the impact of all the pixel values

on the model's risk score. We used this to identify a region of interest (ROI) in the image that contributed the most to the model's output. Our reasoning was that this ROI would be likely to include the cancer if present in the image. We therefore trained a separate resnet101 to classify high-resolution ROIs produced by the Holistic model instead of classifying full images. The approach was successful, as the second resnet performed better than the Holistic one.

However, while the pixel-to-risk score gradient was able to identify relevant parts of an image to some extent, it was very irregular and fluctuated a lot. Also, the sign of the gradient did not convey any useful information, so we were only able to utilize information about its magnitude. In retrospect, this is not surprising, since the modification of a single pixel in isolation is unlikely to have a meaningful impact on the classification of an image. Consequently, the ROIs had to be relatively large to capture the relevant regions consistently. For the present model, we therefore looked for explainable-AI methods that could estimate relevant regions more smoothly and also distinguish between positive and negative contributions to the output of the Holistic model. We considered perturbation-based methods like LIME and Occlusion (8), but concluded that these were too computationally demanding for our purpose. Instead, we went for Layered GradCam (9), which utilizes the fact that CNN models synthesize a hierarchy of increasingly sophisticated features, which can be localized in the image. We ended up using the top convolutional layer, which had a resolution of 31x31. For a given image, this method quantifies the contribution that each location in the given layer gives toward a positive classification by the model. We used this to identify the 7x7 square that in sum contained the largest contribution toward a positive classification and defined this as the model's ROI. For images that got a negative risk score by the model, this procedure tended to choose ROIs outside of the breast, since these gave the least negative risk score contribution. To avoid this, we restricted the candidate ROIs to those with center inside the breast. We considered the more complex method GradCAM++ but decided against it since it has been reported to be numerically unstable (10). This reference also claims that it adds little beyond the standard version, but we have not tested this ourselves.

The second resnet101, called the Focused model, was trained on ROIs from the Holistic model, much like in the previous version. Cross-entropy loss and random flipping, cropping and rotation was used here as well. In Focused model, we were able to utilize the full image resolution without down-sampling. However, it was also advantageous to keep the same physical between-pixel distances, so that the model would not need to generalize over scales. We therefore rescaled all images to a width of 2082 pixels before cropping down the ROI. This had the benefit of keeping a constant physical size of the ROI and a constant number of input pixels for the model (512x512). By contrast, the Attention model of the previous publication had approximately four times as large ROIs (976x976 pixels), which is why we renamed the present one to "Focused". Figure 2 gives a sample mammogram with a screen-detected cancer to the left. For the same mammogram, the middle image shows the gradient point cloud used in the previous model, while the more focused GradCam activation map is shown to the right.

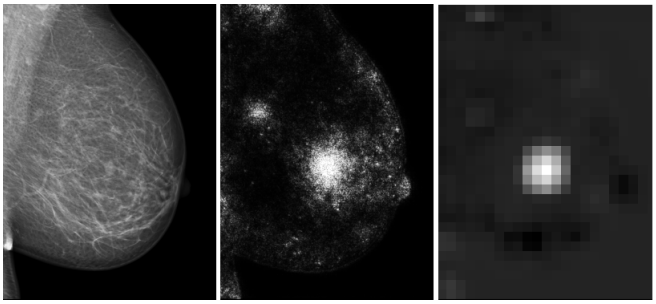

Figure 2: A mammogram, its gradient point cloud and GradCam activation map.

Figure 3 illustrates how the fully trained model-pair works during inference. The lower left part shows Holistic model. It produces a 'preliminary risk score', which served as an auxiliary training task that facilitated the identification of the ROI. The top right part of the figure shows the process where the Focused model evaluates the ROI provided from Holistic model. The top left illustrates how Layered GradCam was applied also to the Focused model during inference, giving a heatmap for ROI, which was embedded into the full mammogram. Unlike the application of this method to the Holistic model, this had the purpose of explaining the model output to the radiologist or clinician in charge of the patient.

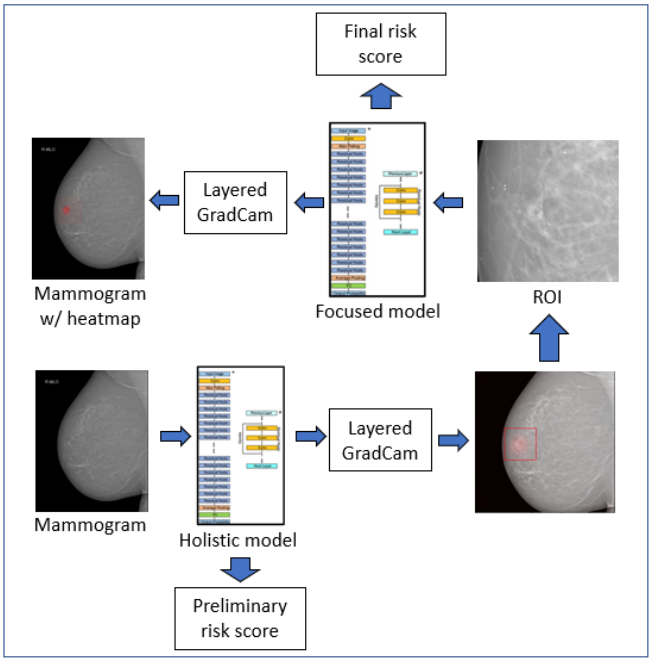

Fig.3: Illustration of the two-stage model.

The illustrations of the Holistic and Focused models in the figure are deliberately 'cartoonish' and not intended to represent the full resnet101 architecture. For details on this, we refer to (7).

The training procedure was split in two phases: First the Holistic model was fully trained end-to-end. In the second phase, the Holistic model was frozen, and served to produce the ROIs for the end-to-end training of the Focused model. In both phases, the image-level cancer status was the only learning signal. While it is often desirable to train a complex model in a single end-to-end process, with the present architecture we found no way of propagating the gradient signals from the Focused model through the ROI identification process and into the Holistic model.

We suspected that the Layered GradCam procedure could fail to locate a cancer that the Holistic model was not able to classify correctly. We therefore implemented an option in Holistic model that produced multiple non-overlapping ROIs, each with the maximum positive cancer-signal in the remaining part of the image. With multiple ROIs, the Focused model evaluated them in sequence and returned the maximum output as its risk score. This was only applied during inference, as the top choice ROI was always used during training.

It is well known that an ensemble of models will often perform better than the average performance of the individual models (11). To this end we use the five models that were trained on different sets of folds. Each of them evaluates the mammogram in question, and the ensemble score is defined as the mean of these risk scores.

So far, we have only discussed risk scores for single mammograms, but real-world applications require examination-level risk scores. For each breast, we treat this in the spirit of the ensemble modelling above, taking the average risk score. For the examination level, we take the maximum of these averages. In the most common case of four images, this amounts to
max(L-CC + L-MLO, R-CC + R-MLO)/2, where L-CC is the ensemble risk score for the left-side CC-image, L-MLO is the ensemble risk score for the left-side MLO image and so on.

The performance of our model was compared with that of a commercially available system. We will refer to this commercial model as "Model X", as our terms of use in this study prevents us from disclosing its name. Model-X also assigned real-valued risk scores to each examination.

We also compared our results with the radiologists' interpretation. Both breasts in each examination had been interpreted by two independent radiologists on a 5-point scale. We defined the single-radiologist score of an examination as the maximum over the left and right scores. For two radiologists, we again used the ensemble approach, taking the average of the two scores for each breast, and used the maximum of these averages as the examination-level radiologist score.

We used the area under the ROC-curve (AUC) for examination classification as our evaluation criterion. AUC has the intuitive interpretation of the probability that a random positive examination be evaluated higher than a random negative one. It is commonly used in the research field and has the favorable property of not being sensitive to class imbalance.

Ethical approval

The research related to human use has been complied with all relevant national regulations, institutional policies and in accordance with the tenets of the Helsinki Declaration and has been approved by the authors' institutional privacy ombudsman and a regional ethics committee for medical research (REC# 2017/2461).

Results

Table 3 gives the AUC performance estimates with confidence intervals for Model X compared to our ensemble model. Due to a technical issue with Model X, this analysis was performed on a restricted part of the data set with 85,054 examinations. Confidence intervals were computed according to (12).

	All cancers (screen-detected and interval)	Screen-detected cancers
Ensemble	0.931 (0.919-0.943)	0.974 (0.968-0.980)
Model X	0.918 (0.905-0.931)	0.959 (0.949-0.968)

Table 3: AUC-performance on the restricted test set from Østfold.

Figure 4 shows the distribution of the examination-level Ensemble risk score for negative cases, screen-detected and interval cancers.

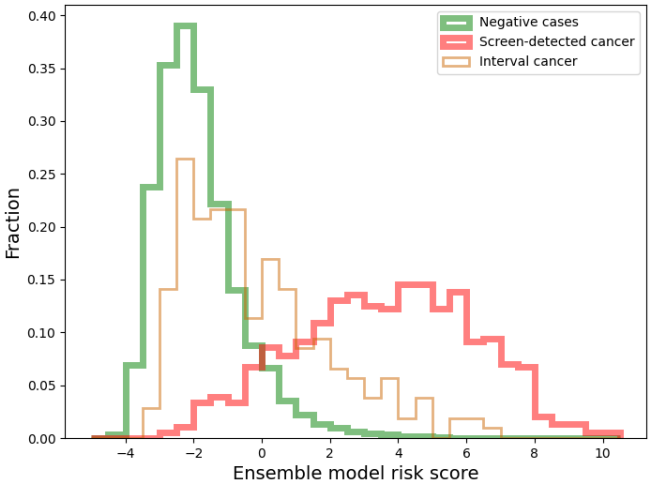

Fig 4: Distributions of examination-level Ensemble risk scores.

The green, red and orange lines represent the distribution for negative cases, screen-detected cancers and interval cancers, respectively. Their areas are each scaled to sum to 1, so they do not represent the fact that negative cases are more numerous. The mean risk score for negatives

is -1.85, for interval cancer -0.14, and for screen-detected cancer 3.62.

Figures 5 and 6 show ROC curves for the Holistic, Focused and Ensemble models for screen-detected cancers and all cancers, respectively. These were based on the set of 124,615 examinations with the four standard views from Østfold.

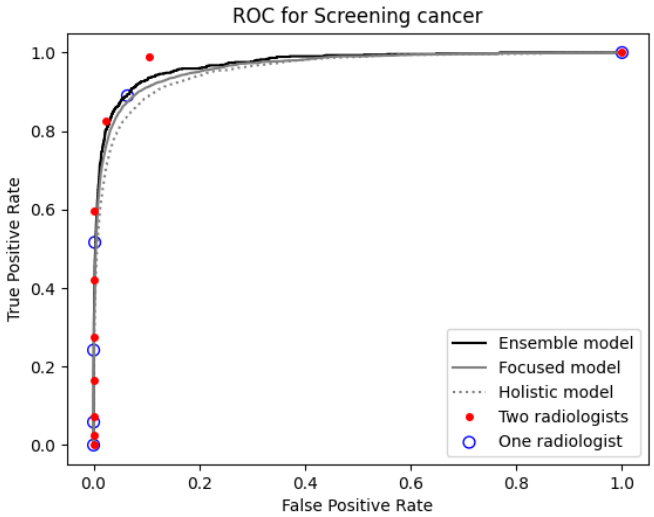

Fig 5: ROC-curves for screen-detected cancer.

Since the radiologist scores are discrete, we show the specific levels as points, identified by blue circles (one radiologist) and red dots (two radiologists).

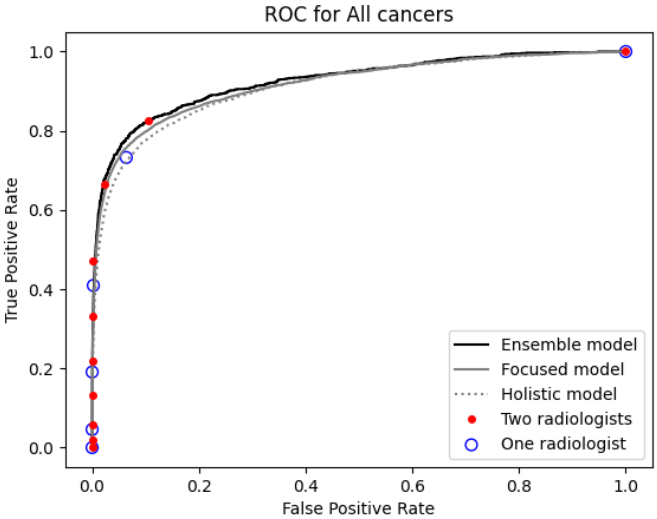

Fig 6: ROC-curves for all cancers (screen-detected and interval cancer).

Our training procedure enabled us to evaluate every examination in the cross-validation set by a Focused model in the ensemble that had not seen the mammograms in question. This enabled us to estimate the Focused model AUC-performance on the full cross-validation data set. This gave 0.962 for screen-detected cancers and 0.917 for all cancers, which was very close to the performance of a single Focused model on the test set (screen-detected cancers: 0.965, all cancers: 0.917). Another test we performed was to increase the number of candidate ROIs from one to three, but the AUC performance was virtually unchanged with less than 0.002 difference.

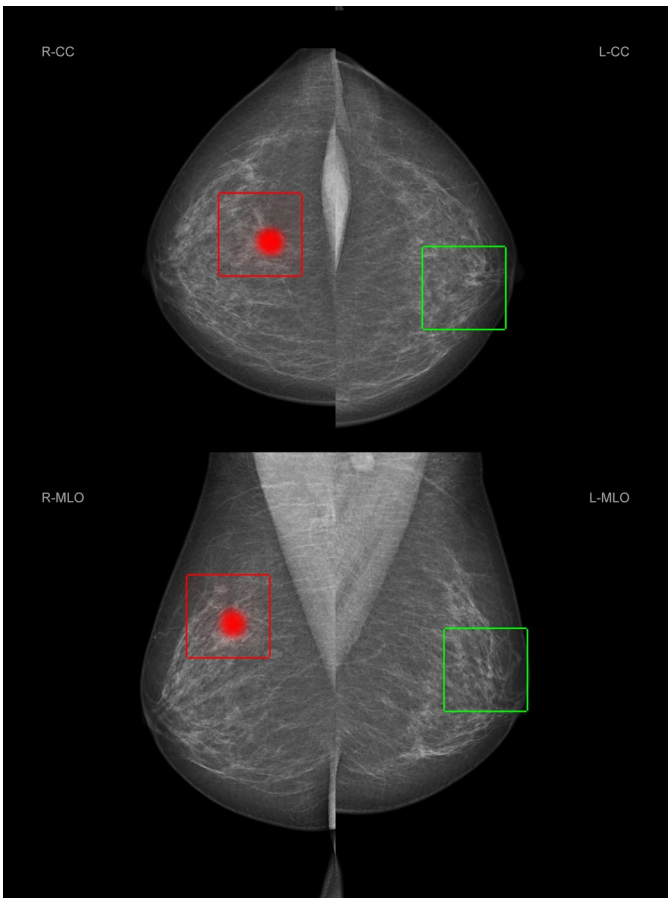

Figure 7: Model output image.

Figure 7 shows an example of the model output for an examination, displayed in a way that is common in mammography. The left breast is shown on the right and vice versa. The top part shows the craniocaudal (CC) views, while bottom part shows the Mediolateral oblique (MLO) views. The squares indicate the ROIs chosen by Holistic model. The red dots indicate a cancer located in the right breast, located by Focused model. The square in the left breast is displayed in green since the Focused model found no signs of cancer there. Incidentally, the Holistic model also gave this mammogram a negative score, but it still identifies the most suspicious area as an ROI for the Focused model to investigate more closely.

Figure 8 shows a false positive case, where the model gave a high score to a negative examination. The right image is the same as the left, with the model's cancer indication added in red.

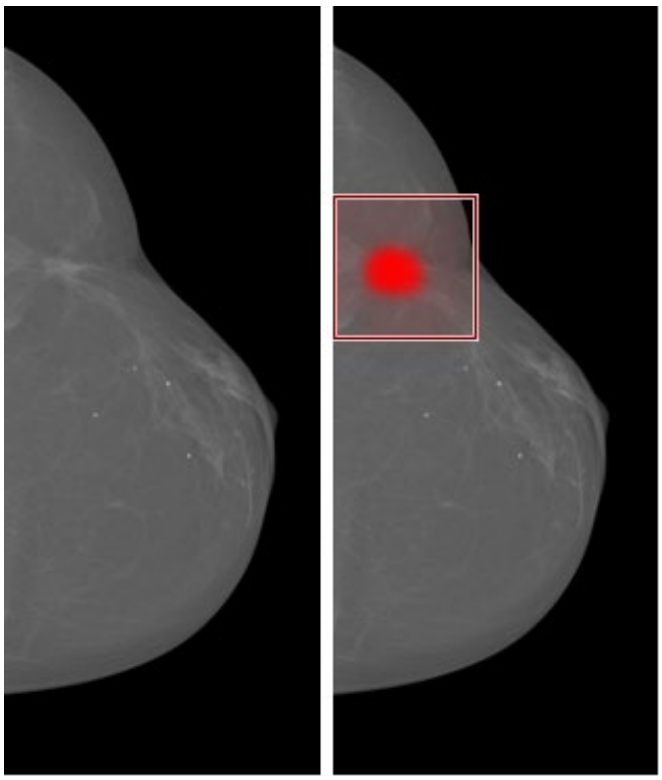

Figure 8: A false positive case without and with annotation by the model.

This breast has a deep scar after surgery, which the model mis-evaluates as cancer. A radiologist would easily recognize this as a scar, partly due to the general shape of the breast. This might be a problem for the Focused model, which looks only at the ROI and loses the bigger picture. We are not very concerned with this type of error, however, since deep scars like this are relatively rare. Also, unlike the model, clinicians or screening personnel are likely to be aware of previous surgeries and may therefore disregard the model's evaluation in cases like this.

Discussion

Over-all our model appears to be successful, as the Ensemble model performed significantly better than a commercial counterpart. The fact that only a subset of the Østfold data was included in the comparison was not likely to favor our model, since the selection was due to a technical issue with Model X.

It is technically possible to calculate ROC curves and AUC values based on the radiologist scores, but this would not do them full justice. The resulting ROC curve would consist of straight lines between the discrete points in the figures, which would give lower AUC values compared to the curved lines for the models. It makes more sense to compare the points to the ROC curves of the models directly. When a radiologist's point is below a model's ROC curve, it means that if we tune the decision limit so that the model's true positive rate (TPR) be equal to the radiologist's, the model's false positive rate (FPR) becomes lower. Alternatively, tuning the model's FPR to the radiologist gives a higher TPR. When the point is above the curve, tuning the TPR gives a higher FPR than the radiologist and tuning the FPR gives a lower TPR. Figure 5 shows that for screen-detected cancer, the Ensemble model is comparable to a single radiologist since the blue circles are located on the ROC curve, while two radiologists perform better, as indicated by the red dots.

Using screen-detected cancers as the gold standard is likely to favor the radiologists over our model since an elevated score by at least one radiologist is required for a cancer case to be labeled 'screen-detected cancer'. Some of the interval cancers would have been possible to detect at the given examination, as demonstrated by the fact that the model gives them an elevated risk score. The two radiologist scores differ quite often, and a substantial number of the screening-detected cancers were flagged by only one of them (13). It is therefore likely that some actual cancers will have been flagged by neither, and some of these would turn up as interval cancers. When the model gives these a high score, it is unreasonable to insist that they constitute false positives and reduce the model's AUC for flagging them. Therefore, the analysis of all cancers in Figure 6 may give a more balanced picture. Here we see that the two radiologists follow the Ensemble model curve closely, while one radiologist performs between the Holistic and Focused models. This indicates that one of the radiologists may safely be replaced by the Ensemble model.

This approach coincides with the one taken in the on-going Swedish study mentioned in the Introduction (4), where one of the two radiologists was replaced with a commercial AI product and used as a decision support in about 90% of the examinations, and used as a decision support for both radiologists in the 10% of the cases with the highest score. Their intermediate safety analysis after the inclusion of 80,000 women showed a higher cancer detection rate for the women that were randomized to the AI-supported study arm, without an increase in false positives. This is in line with our present results, as shown in Figure 2. Interval cancers are not yet analyzed in the study from Sweden.

Compared to radiologist scorings, AI-models in general have the advantage of a continuous outputs, which makes it possible to adjust the desired fraction of cases that should be investigated further.

The improvements from the Holistic to the Focused and Ensemble models, as shown in Figures 5 and 6, may seem modest. There is, however, a strong effect of diminishing returns, as each additional cancer case is harder to find.

While the Holistic model separated positive from negative images rather well, it did an even better job at identifying the relevant ROIs. This means that the explainable-AI algorithm Layered GradCam applied to Holistic model succeeded in locating cancers that the model itself could not reliably classify as such. This conclusion was strengthened by the fact that multiple ROIs did not improve the performance in a meaningful way. Hence, our two-stage

procedure makes it possible to train a high-quality model without the need for time-consuming human pixel-level annotation of lesions.

We believe the primary reason why the Focused model performed better than the Holistic one to be that it was able to utilize full (or close to full) image resolution. However, it will also have benefited from a more favorable signal-to-noise ratio, in that a larger part of its input field contained signs of cancers. This is supported by the observation that it required fewer training epochs to converge (details not included).

With machine learning, one is often worried that a model may perform worse on independent data. It is therefore very encouraging that our model performs no worse on the Østfold test set as on the cross-validation on the development set.

It should be noted, however, that these analyses were restricted to images produced with Siemens equipment, and our model may be less robust for images from other manufacturers.

We use the same Layered GradCam XAI-method also for communicating the model output to the radiologist or clinician responsible for the patient. Radiologists are familiar with the format presented in Figure 7 and has been useful in human expert validations of the model.

Conflict of interest

Authors state no conflict of interest.

References

1. Lauby-Secretan B, others. Breast-Cancer Screening — Viewpoint of the IARC Working Group. N Engl J Med. 2015;372:2353-2358.

2. Hofvind S, others. The Norwegian Breast Cancer Screening Program, 1996-2016: Celebrating 20 years of organised mammographic screening. 2017.

3. McKinney SM, others. International evaluation of an AI system for breast cancer screening. Nature. 2020;577(7788):89–94.

4. Lång K, Josefsson V, Larsson AM, Larsson S, Högberg C, Sartor H, et al. Artificial intelligence-supported screen reading versus standard double reading in the Mammography Screening with Artificial Intelligence trial (MASAI): a clinical safety analysis of a randomised, controlled, non-inferiority, single-blinded, screening accuracy study. Lancet Oncol. 2023 Aug 1;24(8):936–44.

5. Dahl F, Holden M, Brautaset O, Eikvil L. A mammography classification model trained from image labels only. Proc North Lights Deep Learn Workshop [Internet]. 2022 Mar 28 [cited 2022 Mar 30];3. Available from: https://septentrio.uit.no/index.php/nldl/article/view/6244

6. Zhang Y, Yang Y. Cross-validation for selecting a model selection procedure. J Econom. 2015 Jul;187(1):95–112.

7. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) [Internet]. 2016 [cited 2023 Oct 22]. p. 770–8. Available from: https://ieeexplore.ieee.org/document/7780459

8. Gulum MA, others. A Review of Explainable Deep Learning Cancer Detection Models in Medical Imaging. Appl Sci. 2021;11(10).

9. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. Int J Comput Vis. 2020 Feb;128(2):336–59.

10. Lerma M, Lucas M. Grad-CAM++ is Equivalent to Grad-CAM With Positive Gradients [Internet]. arXiv; 2022 [cited 2023 Oct 3]. Available from: http://arxiv.org/abs/2205.10838

11. Ganaie MA, others. Ensemble deep learning: A review. 2021.

12. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982 Apr;143(1):29–36.

13. Martiniussen MA, Sagstad S, Larsen M, Larsen ASF, Hovda T, Lee CI, et al. Screen-detected and interval breast cancer after concordant and discordant interpretations in a population based screening program using independent double reading. Eur Radiol. 2022 Sep 1;32(9):5974–85.