Reducing False-Positive Biopsies using Deep Neural Networks that Utilize both Local and Global Image Context of Screening Mammograms

Nan Wu[1] · Zhe Huang[1] · Yiqiu Shen[1] · Jungkyu Park[1] · Jason Phang[1] · Taro Makino[1] · S. Gene Kim[2,3,4] · Kyunghyun Cho[1,5,6] · Laura Heacock[2] · Linda Moy[2,3] · Krzysztof J. Geras[1,2,4]

Abstract

Breast cancer is the most common cancer in women, and hundreds of thousands of unnecessary biopsies are done around the world at a tremendous cost. It is crucial to reduce the rate of biopsies that turn out to be benign tissue. In this study, we build deep neural networks (DNNs) to classify biopsied lesions as being either malignant or benign, with the goal of using these networks as second readers serving radiologists to further reduce the number of false-positive findings. We enhance the performance of DNNs that are trained to learn from small image patches by integrating global context provided in the form of saliency maps learned from the entire image into their reasoning, similar to how radiologists consider global context when evaluating areas of interest. Our experiments are conducted on a dataset of 229,426 screening mammography examinations from 141,473 patients. We achieve an AUC of 0.8 on a test set consisting of 464 benign and 136 malignant lesions.

Keywords Breast cancer · Deep neural networks · Screening mammography · Global context · Local patterns

Introduction

Breast cancer is the most common cancer in women worldwide, after skin cancers and about 42,170 women will die from breast cancer in the United States for 2020, according to The American Cancer Society's estimate [1]. Screening mammography, a low-dose X-ray examination, is typically used for early detection of breast cancer. The United States Preventive Services Task Force suggests women urdergo such examinations every two years if they are 50 to 74 years old and are at average risk for breast cancer [2]. Although multiple studies have demonstrated that screening mammography reduces breast cancer mortality [2–5], performance benchmarks demonstrate that 10% of the performed examinations are recalled for additional imaging, and approximately 80% of biopsies subsequently performed are benign [6]. The yearly national cost of breast care caused by the false-positive mammograms is estimated to be several billion dollars [7, 8] and for women with a false-positive diagnosis, their mean cost of breast care is even higher than the cost of breast care services for women with cancer [9]. It is therefore an important task to reduce the recall and biopsy rates so that to decrease patients' anxiety and reduce healthcare costs while still maintaining optimal cancer detection rates, according to relevant guidelines [5].

Traditional computer-aided detection (CAD) tools for mammography neither detected more breast cancers nor decreased the recall rates for additional imaging [10, 11]. Early studies used deep neural networks (DNNs) to assist radiologists intrepreting screening mammograms by making predictions for cancer of each breast [12–18]. This task is frequently considered in the literature. It can be viewed

✉ Nan Wu
nw1045@nyu.edu

1   Center for Data Science, New York University,
    New York City, USA

2   Department of Radiology, New York University School
    of Medicine, New York City, USA

3   Perlmutter Cancer Center, NYU Langone Health,
    New York City, USA

4   Center for Advanced Imaging Innovation and Research, NYU
    Langone Health, New York City, USA

5   Department of Computer Science, Courant Institute
    of Mathematical Sciences, New York University,
    New York City, USA

6   CIFAR Associate Fellow, New York City, USA

as breast-level classification, and models developed accordingly have shown comparable performance to radiologists [15–17]. However, these models suffer from performance degradation when evaluated on a population only containing examinations which lead to biopsies, without healthy breasts as negative cases [17]. Meanwhile, models built for the breast-level classification task cannot provide independent risk estimations for multiple areas of interests appearing in the same breast. It is common to encounter cases with multiple findings [19]. For example, multiple bilateral circumscribed breast masses are detected in approximately 1.7% of routine screening mammograms [20]. In the NYU Breast Cancer Screening Dataset [21], a representative sample of screening mammograms from 2010 to 2018, there are 7.45% images with more than one annotated lesions, and 25.75% of these images have lesions of different categories. Some examples are shown in Fig. 1. In light of this, the previously proposed models for breast-level classification are difficult to use for the goal of reducing unnecessary biopsies.

Besides breast-level malignancy classification, deep learning methods have also been used to identify high-risk lesions [22–25]. Some of these works can provide risk estimation across regions of the breast, but usually only consider information in a small local region [22, 25]. The majority of the existing works often utilize object detection models such as Mask-RCNN [26], which neither explicitly utilize fine details nor consider global context. In contrast, radiologists often consider global context factors to make their diagnoses [27, 28]. These global context factors include the

mammographic breast density, i.e. the global amount of fibroglandular tissue, and the associated parenchymal and nodular patterns of the breasts [27]. Dense fibroglandular breast tissue is a known risk factor for breast cancer [29]. Other global context factors include the distribution of microcalcifications in the tissue adjacent to an index lesion, or throughout the breast. These global findings often affect radiologists' level of suspicion for a particular lesion. Within deep learning methodology, these scenarios could be viewed as utilizing global image context for classifying a patch of an image. This motivates investigating whether global context is as important for neural networks as it is for human experts.

In this study, we consider lesion-level classification, and design models to directly distinguish biopsy-confirmed lesions as being either benign or malignant. With this strategy, we enable the models to make accurate lesion-wise predictions. To show that deep learning approaches can benefit from utilizing global image context in classifying local findings on mammograms, we first train DNNs with cropped image patches to enable the learning of local information from a specific region, then integrate the extracted local information with the global context. The global context is provided in the form of saliency maps (Fig. 2) extracted by a model classifying the entire image. Here we use Globally-Aware Multiple Instance Classifier [16] as the model to provide such saliency maps. In addition, we evaluate the models' performance on a challenging population which consists only of difficult to diagnose cases for which the radiologist requested a biopsy. This further differentiates our work from previous works [14–17, 30] and makes our results not directly comparable to theirs. This is because these methods were developed and evaluated for the screening population, which contains a lot of negative cases not requiring biopsy, which can inflate their evaluation metrics [17].

Our results show that DNNs trained with image patches can effectively decrease the number of unnecessary biopsies, and that this ability can be further improved by utilizing global image context. Our best model is able to distinguish between benign and malignant findings on a test set of 600 lesions, achieving an area under the receiver operating characteristic curve (AUC) of 0.799±0.002. If the model is utilized to assist in reading mammograms, about 2% unnecessary biopsies could be avoided, in addition to cases that radiologists can easily exclude as benign or normal and not needing additional imaging, while catching all malignancies. It reduces 15% more unnecessary biopsies than the model using only local information, when missing 2% malignancies. It is worth noting that these performance measurements are computed on the population from which we exclude benign cases that radiologists can discount confidently by reading mammograms or other imaging examinations. Overall, our results strongly suggest that the proposed strategy can be considered as a viable and valuable enhancer for deep
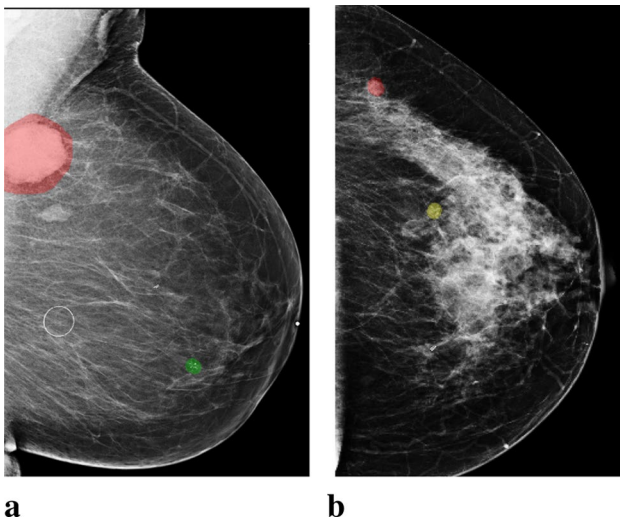
Fig. 1 Images with both malignant and benign lesions. **a**. An image of the left breast from mediolateral oblique view (L-MLO). The breast has two lesions confirmed by biopsy, one as malignant (annotated with red), and the other as benign (annotated with green). **b**. An L-MLO mammogram image from another patient. There are two lesions on the image, one as malignant (annotated with red), and the other as high-risk benign (annotated with yellow)

Fig. 2 An example of saliency maps. From left to right: a mammogram image of a right breast from craniocaudal view (R-CC) with an annotated malignant lesion, a saliency map indicating suspicious regions for benign lesions, a saliency map indicating suspicious regions for malignant lesions
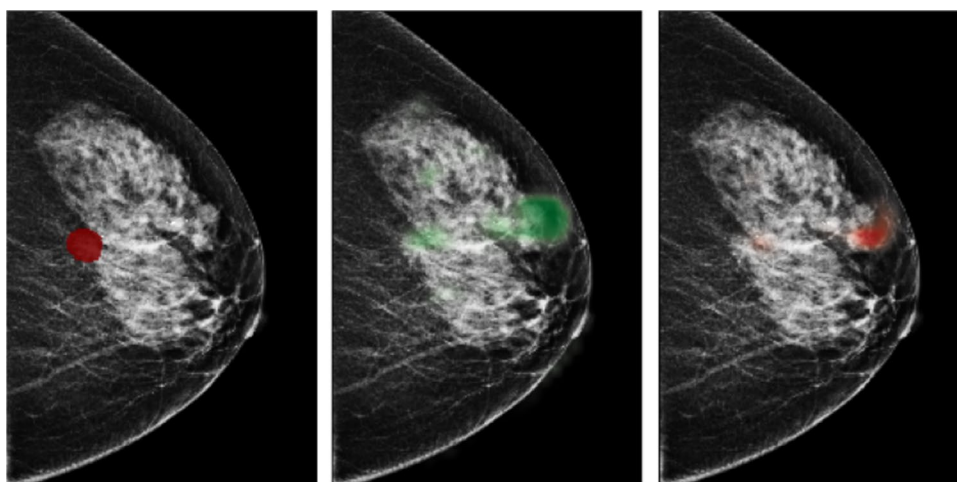
learning methods in reducing unnecessary biopsies based on screening mammography.

Materials and Methods

This retrospective study was approved by our IRB and is compliant with the Health Insurance Portability and Accountability Act. Informed consent was waived.

Data

We utilize a dataset consisting of 229,426 digital screening mammoramphy examinations (1,001,093 images) from 141,473 unique patients screened between 2010 and 2017 [21]. Each examination has four standard views and the resolution of images is approximately 2000×3000 pixels. We asked fellowship-trained breast imagers to annotate both benign lesions (e.g. cyst, fibroadenoma, fibrocystic change) and malignant lesions (e.g. IDC, ILC, DCIS), on the pixel-level. In the entire dataset, there are 8842 lesions from 8080 images with diagnosis confirmed by biopsy, which reveals the fact that a single breast can contain multiple lesions of differing types. The dataset is divided into disjoint training (80%), validation (10%) and test (10%) sets. Detailed statistics of training, validation and test sets are in Table 1.

The Proposed Method

Lesions in mammograms vary in size and shape, so if we crop these regions entirely and resize them to the same size to use them as inputs to standard deep neural networks, we will introduce information distortion and lose the fine details of the lesions. Therefore, we start by learning features of a number of image patches that are cropped from regions overlapping with one of the lesions, and then aggregate information from all patches to render a prediction for that lesion.

To extract information from image patches, we train a deep convolutional neural network (DCNN) to classify image patches of 256×256 pixels as one of the four classes: "malignant", "benign", "outside" and "negative." Malignant and benign patches are cropped from windows that overlap with the segmentation of a malignant or benign lesion. Besides cropping image patches that overlap with the annotations, we sample patches that have no overlap with any lesion ("outside"), as well as patches from breasts without records of biopsy ("negative"). The inclusion of these additional data is intended to regularize the model similarly to data augmentation. Examples of patches from each class are shown in Fig. 3.

This DCNN is used to produce representations of local information. It is denoted as f_{loc} and is shown in Fig. 4a. We use DenseNet-161 [31] as its architecture. We add an additional fully connected layer with 32 neurons between the global average pooling layer and the classification layer to obtain concise representations of the patch. The additional layer results in the feature vector $\mathbf{h} \in \mathbb{R}^{32}$ for the patch, which we use as the representation for the local information extracted by f_{loc}.

To further incorporate global image context and curate the local information extracted by the DCNN, we train an "aggregation network" with inputs formed by aggregating

Table 1 Number of biopsy-confirmed lesions and number of mammogram images presenting no lesions, benign lesions, and malignant lesions in the training, validation and test set

|	images			lesions	
	negative	benign	malignant	benign	malignant
training	808,730	5,188	1,648	5,602	1,790
validation	123,130	687	110	722	128
test	60,959	432	116	464	136
overall	992,819	6,307	1,874	6,788	2,054

Fig. 3 Examples of image patches along with the mammogram images from which they come. **a**. "malignant" patches, which overlap only with malignant findings (marked with red); **b**. "benign" patches, which overlap only with benign findings (marked with yellow or green); **c**. "outside" patches, which are from regions outside the annotated lesions; **d**. "negative" patches, which are from images without any biopsied findings
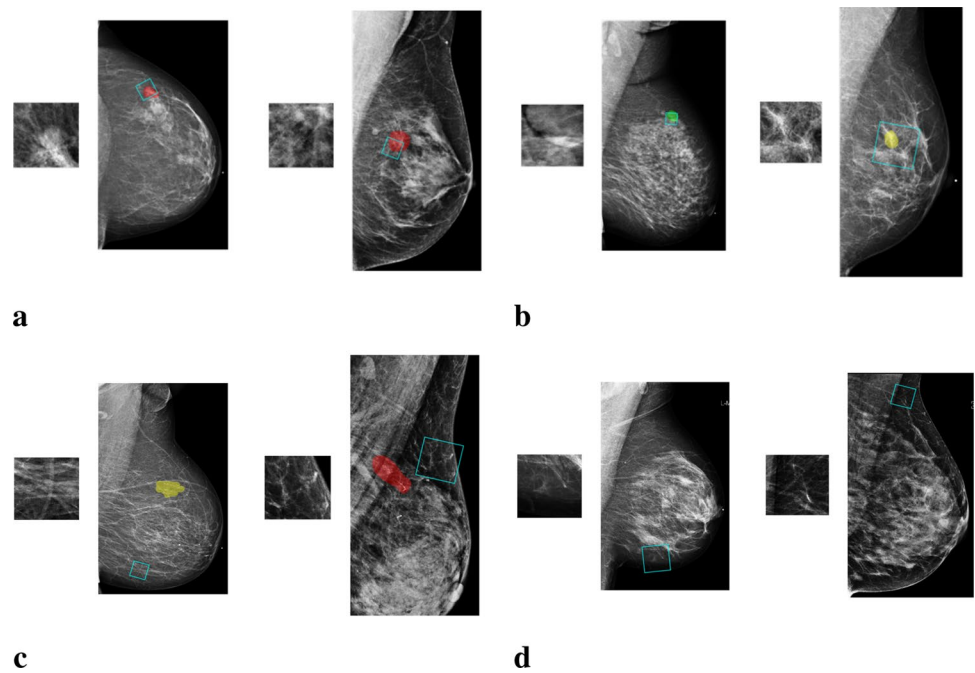
maps containing information relative to the patch and the image it is cropped from, as illustrated in Fig. 4b. This aggregation network is a shallow convolutional network, denoted as f_{agg}. It consists of two convolutional layers, each with 32 3×3 convolutional filters, a global average pooling layer and, finally, a classification layer. We apply batch normalization and the ReLU activation function prior to each convolutional layer. This network is trained for the same patch classification task. The maps formed as inputs to the aggregation network are described in the following paragraphs.

The first type of maps is saliency maps which represent global context. We generate the saliency maps by training a network on full-resolution mammography images to predict the presence of benign and malignant lesions in the breast. We refer to this network as the "context network." We use Globally Aware Multiple Instance Classifier [16, 30] as the context network, which is explicitly designed to provide interpretability by highlighting the most informative regions of the input images. To be more precise, the feature maps obtained after the last residual block of the context network are transformed by a 1×1 convolutional layer

Fig. 4 Illustration of the proposed method. Left: a deep convolutional neural network which takes image patches of 256× 256 pixels as inputs, denoted as f_{loc}. Right: the aggregation network, f_{agg}, that takes the concatenation of three types of maps as inputs: 1) location indicator map, **I**, in gray, generated by downscaling the binary mask indicating the cropping window's location, 2) saliency maps, $\mathbf{S}_m$ and $\mathbf{S}_b$, in red and green, generated by the context network, based on Globally Aware Multiple Instance Classifier [16], 3) embedding map, **E**, in blue, formed by the representation **h**, produced by f_{loc}
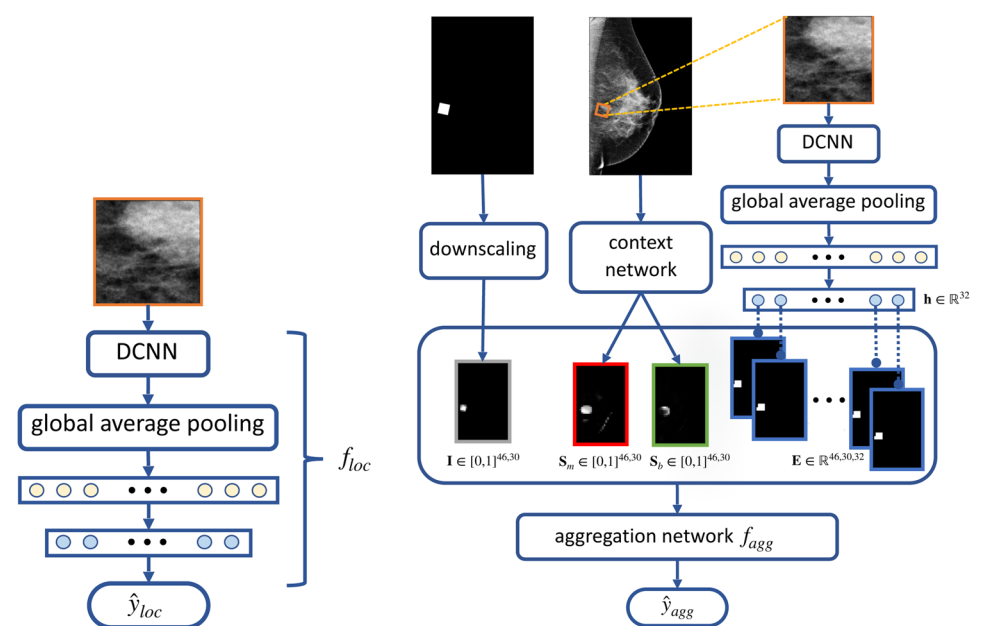
with sigmoid activation into two saliency maps, denoted as $\mathbf{S}_m \in [0,1]^{46,30}$ and $\mathbf{S}_b \in [0,1]^{46,30}$. Each pixel in the saliency map corresponds to a region in the full image, and its element denotes a score indicating the contribution of this region towards classifying the input image as containing malignant or benign lesions. A pair of saliency maps for an image is shown in Fig. 2.

Another type of maps are location indicator maps. Given a patch, this map indicates its cropping window's location on the mammogram, but is downscaled to be the same size as the saliency maps. The location indicator map is denoted as $\mathbf{I} \in [0,1]^{46,30}$. Same with the saliency maps, each pixel on the location indicator map corresponds to a region in the full image and the value of this pixel reflects how much the region is covered by the patch.

The last type of maps, called embedding maps, are formed utilizing the representation $\mathbf{h}$ generated by f_{loc}. To construct this map, for each $\mathbf{h}_k \in \mathbf{h}$, we take a copy of the location indicator map of the patch $\mathbf{I}$ and replace its nonzero elements by $\mathbf{h}_k$. We denote the obtained map by $\mathbf{E}_k$. We concatenate $\mathbf{E}_k$'s to form the full embedding map $\mathbf{E} \in \mathbb{R}^{46,30,32}$. The embedding maps contain information learned by f_{loc} specific to the fine details in the image patch.

These maps can be concatenated along the last dimension and served as inputs to the aggregation network, denoted as $\mathbf{X} \in \mathbb{R}^{46,30,M}$ where M is the number of maps. For example, when using both embedding maps and saliency maps as inputs, M would be 34.

Model Training

We first train f_{loc} that takes image patches as inputs, followed by the aggregation network f_{agg}, which takes the concatenated maps, $\mathbf{X}$, as inputs. We use 20, 35, 5000, 4945 patches for malignant, benign, outside and negative patch classes in each training epoch. For data augmentation, we use random rotations (-30 to 30 degrees), and random sizes (128×128 to 384×384 pixels) when setting the cropping window to obtain the patch.

In order to address the extreme class imbalance, we use weighted cross-entropy as the training loss. The class weight for each patch class is set as inverse to the ratio of patches from this class among all patches used in each epoch. Therefore, losses on incorrect predictions of "malignant" and "benign" patches are appropriately up-weighted. For both f_{loc} and f_{agg}, we adopt the same configuration while using 256×256 image patches as inputs for f_{loc}, and the concatenated maps, $\mathbf{X}$, as inputs for f_{agg}, in which the embedding maps is produced by the best performing f_{loc} and saved to be used.

We minimize the training loss with the Adam optimizer [32], setting the batch size to 25 for training f_{loc}, and 100 for f_{agg}. We initialize weights in f_{loc} of all layers

except the last two fully connected layers with weights from DenseNet-161 [31] pretrained on the ImageNet ILS-VRC-2012 dataset [33], then fine-tune the entire network. We randomly initialized the weights of f_{agg}. We optimize the hyperparameters using random search [34]. Specifically, we search for the learning rate on the logarithmic scale in $[10^{-6}, 10^{-4}]$ for f_{loc}, and in $[10^{-5}, 10^{-3}]$ for f_{agg}. Early stopping is performed if we observed that the AUC on the validation set has not increased for ten epochs. We implement the models in PyTorch [35], and use NVIDIA Tesla V100 GPUs for model training and inference.

Model Evaluation

During training, we consider patches from mammograms with and without lesions, and perform multi-class classification over four classes: malignant, benign, outside and negative. In the validation and test phases, we only consider patches from images with lesions, and transform the patch-level predictions into a malignancy prediction for each lesion in the images. To get a prediction for a lesion, as shown in Fig. 5, we crop 100 patches that overlap with the segmentation of the lesion. The size of the cropping window varies from 128×128 to 384×384 pixels, which is the same range we used for data augmentation. After cropping, each patch is resized to 256×256 pixels, and we use it as input to f_{loc} to produce a feature vector. Then, we apply f_{agg} on the concatenated maps, including the embedding maps transformed by the feature vector and we get its prediction, each as four scores for the four patch classes. For each patch, we normalize the scores for malignant and benign patch classes so that they sum to one. Finally, we average the 100 normalized scores of the 100 sampled patches to obtain a prediction for the lesion. Based on these estimated probabilities, we compute the AUC that the model achieves in classifying the lesions as malignant or benign. We use the AUC computed on the 850 lesions from the validation set for model selection, and report the AUC computed for the 600 lesions from the test set.

Results

Model Performance

We report the model's performance on the test set consisting of 600 lesions which are present on 534 images from 260 patients. There are 44 images containing more than one lesion, and 14 images have both benign and malignant lesions. We emphasize that if we were to use deep learning methods that provide only breast-level risk estimation, it would be impossible to tell which lesion is the one with higher risk of malignancy.

Fig. 5 Multiple patches are sampled to obtain a prediction for one lesion. In each black box, we present a biopsied finding from the test set and ten out of 100 patches used by the model to make a prediction for the finding. Lesions are marked in red (malignant) or green (benign) and the cropping windows of the patches are marked by blue boxes on the image. Cropped patches are shown on the right
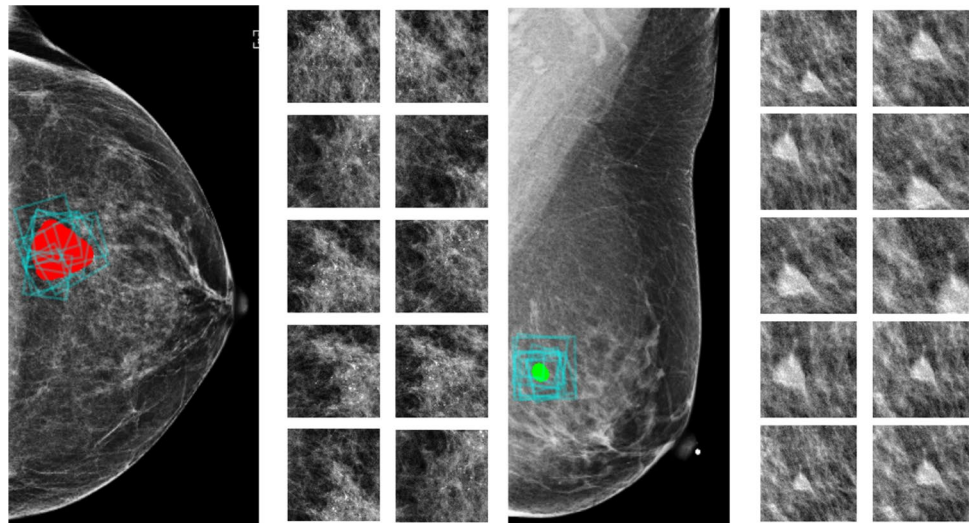
Both model components, f_{loc}, which takes patches as inputs, and the aggregation network f_{agg}, are selected according to their performance on the validation set. The best performing aggregation network using embedding maps and saliency maps achieved an AUC of 0.799±0.002 on the test set. In Table 2, we include more results on the portion of unnecessary biopsies that can be avoided while missing a given portion of malignancies when using the model's prediction as a second reader to assist radiologists. It can help to further reduce 1.7% unnecessary biopsies in addition to cases that radiologists can easily exclude as benign or normal and not needing additional imaging, while catching all malignancies. According to the estimated yearly cost related with unnecessary biopsies [7, 8], it can be translated to saving more than a million dollars each year in the US for breast care. If reducing biopsies is prioritized further, up to 13.5% could be avoided while only missing 1% of malignancies and up to 23.1% could be avoided while only missing 2% of malignancies.

Ablation Experiments

We conduct the following experiments to justify the choice of model architecture, to verify impact of transfer learning,

and to elaborate the importance of utilizing both local fine details and global image context in identifying malignant lesions.

Architecture Search. To choose the architecture for f_{loc}, we considered a number of ResNet [36] and DenseNet [31] variants. These architectures use skip connections, which improve information flow between layers, and allow for effective training of very deep networks. They both achieved strong results across a wide range of image classification tasks [37–39]. The specific ResNet and DenseNet variants we experimented with are: ResNetV2-50, ResNetV2-101, ResNetV2-152, DenseNet-121, DenseNet-161, and DenseNet-169. We compared the performance of f_{loc} when being parameterized as the above architectures. The results are shown in Fig. 6. In this experiment, we did not consider global context and used the predictions made by f_{loc} for each lesion.
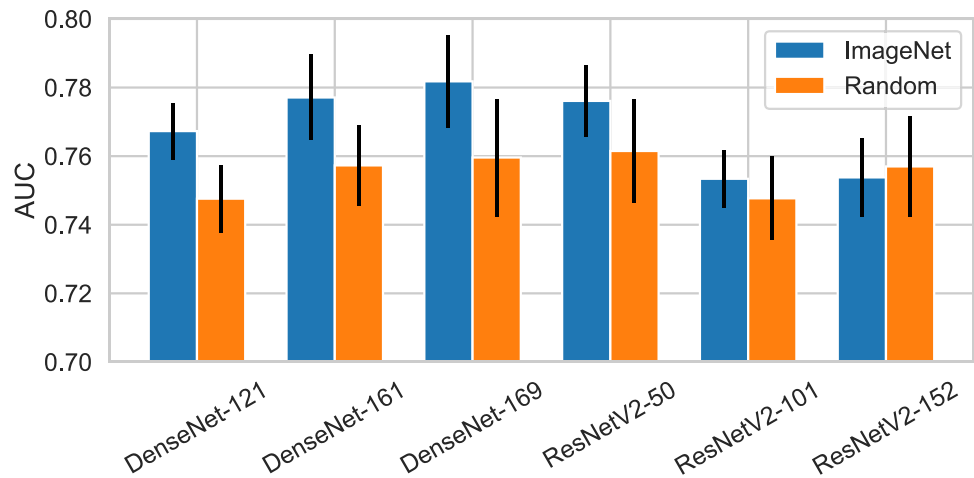
Transfer Learning. Transfer learning by pretraining the network on a different task is widely adopted to improve neural networks' performance. We experiment with initializing f_{loc} with weights from networks pretrained on the ImageNet ILSVRC-2012 dataset [33], and compare it to initializing the weights randomly using He initialization [40]. Since images from ImageNet are RGB while mammograms

Table 2 True negative rate (TNR) achieved by our model when its false-negative rate (FNR) are 0.01, 0.02, 0.03 and 0.05 as we vary prediction threshold for assigning observations to a positive class

indicating malignancy. The 95% confidence interval of the estimated TNR and the clinical meaning are also provided

clinical meaning	FNR	TNR	95% CI of TNR
1.7% unnecessary biopsies we can help to avoid while missing no malignancies	0.00	0.017	[−0.005, 0.040]
13.5% unnecessary biopsies we can help to avoid while missing 1% malignancies	0.01	0.135	[0.039, 0.232]
23.1% unnecessary biopsies we can help to avoid while missing 2% malignancies	0.02	0.231	[0.190, 0.273]
27.6% unnecessary biopsies we can help to avoid while missing 3% malignancies	0.03	0.276	[0.240, 0.312]
43.5% unnecessary biopsies we can help to avoid while missing 5% malignancies	0.05	0.435	[0.407, 0.562]

Fig. 6 Test performance in classifying biopsied findings, achieved by the DCNN, f_{loc}, when using different architectures and weights initialization strategies

are grayscale, we duplicated each patch across the three channels. The AUCs achieved by f_{loc} with or without using transfer learning are presented in Fig. 6. Without transfer learning, ResNetV2-50 achieved the highest AUC of 0.762±0.015, while DenseNet-121 achieved the lowest AUC of 0.748±0.098. When we applied transfer learning, the performance is improved for most of the architectures except ResNetV2-152, and DenseNet-169 becomes the best performer with 0.782±0.014 AUC. We conclude from these results that transfer learning from the ImageNet dataset [33] clearly improves our results, even though the natural image domain and the mammography image domain are so different.

Importance of Global Context. To assess the importance of global context in classifying lesions localized to small regions of the image, we performed further ablation experiments. We trained networks using different combinations of saliency maps, location indicator map and embedding maps as inputs. Selected maps are concatenated along the last dimension and used by the aggregation network. Table 3 presents the results when using all possible combinations. Since ImageNet-pretrained DenseNet-161 as f_{loc} achieved the highest AUC on the validation set, we used it in this set of experiments.

Table 3 Test performance of the aggregation network when using different information combinations as inputs. Models utilizing both local and global information achieved better performance than the counterparts using single type of maps

|	AUC
location indicator maps	0.474 ± 0.031
embedding maps	0.778 ± 0.002
saliency maps	0.695 ± 0.003
location indicator maps + embedding maps	0.777 ± 0.002
location indicator maps + saliency maps	0.721 ± 0.011
embedding maps + saliency maps	**0.799 ± 0.002**
location indicator maps + saliency maps + embedding maps	0.797 ± 0.001

As expected, for the task of classifying biopsy-confirmed lesions, most of the predictive power comes from local features: the aggregation network trained with only embedding maps achieved an AUC of 0.778±0.002. In comparison, the network trained only with saliency maps achieved an AUC of 0.695±0.003, indicating that global context alone was not highly predictive. When we introduced location indicator maps together with saliency maps into the network, the AUC increased to 0.721±0.011. We observed that using location indicator maps and embedding maps together does not improve performance. This is unsurprising since embedding maps contain the same location information conveyed by the location indicator maps. Finally, networks using the combination of embedding maps and saliency maps achieved an AUC of 0.799±0.002, which is the highest among all combinations. The fact that combining local features with global context outperformed each of them in isolation confirms the importance of the global image context in classifying lesions localized to small regions of the mammogram.

To further evaluate the impact of the global context in our task, we investigate the relation between the true negative rate (indicating biopsies that can help to avoid) and the false-negative rate (indicating missed malignancies). Specifically, we compare the aggregation network using only the ensemble maps and the aggregation network using both the ensemble maps and the saliency maps. Figure 7 visualizes this relationship. We considered scenarios when there are less than 5% malignancies missed. For all considered false-negative rates utilizing the saliency maps resulted in lower true negative rate.

Discussion

Regular screening mammography is widely acknowledged to be the best way to detect breast cancer early. However, mammogram-based diagnosis performed by radiologists suffers

Fig. 7 True negative rate (TNR) and false negative rate (FNR) achieved by the aggregation network using ensemble maps or using both ensemble maps and saliency maps as we vary the prediction threshold for assigning observations to a given class
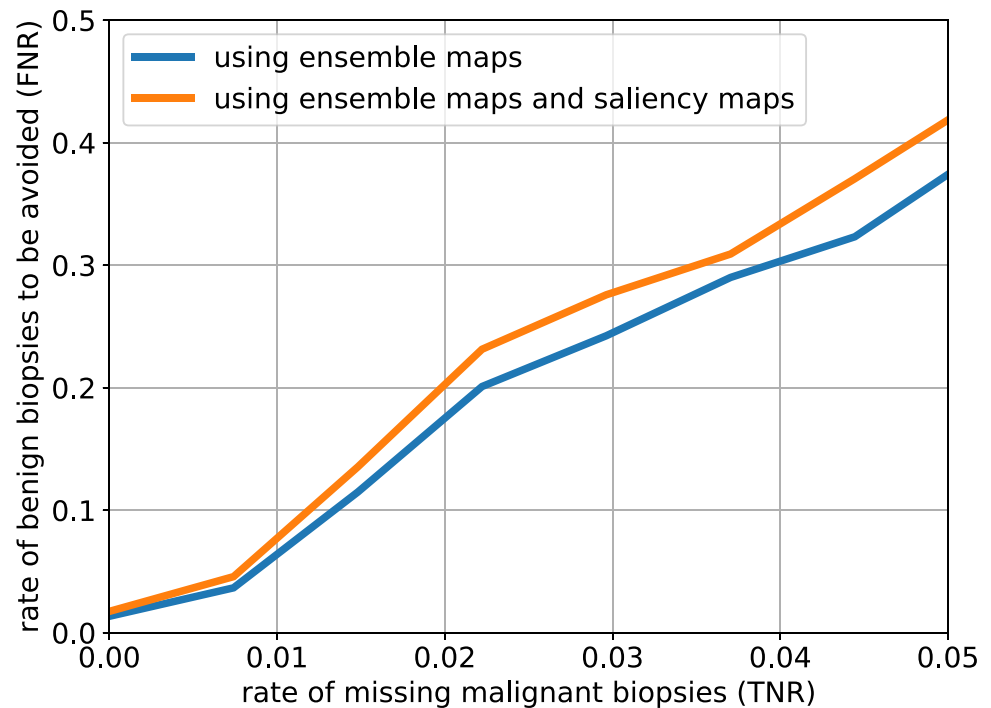
from a high false positive rate, resulting in both unnecessary imaging and tissue biopsies. Developing deep learning technologies to assist breast cancer screening is promising, but previous works in the literature rarely focused on reducing unnecessary biopsies. Besides achieving radiologist-level performance at detecting breast cancer in mammograms, deep learning models are expected to play a more important role in distinguishing whether a given lesion is malignant or benign. This distinction is highly beneficial for the case of suspicious-appearing but ultimately benign findings that result in unnecessary biopsies by the radiologist.

In this study, we presented a method to combine local features in small image patches with global context in high-resolution mammogram images. We showed that it is necessary to consider both fine details in a small region and the global image context to improve deep learning models' performance when classifying localized lesions on the high-resolution images, while previous works usually consider only image patches or downscaled mammogram images [12, 18, 41]. Our resulting deep learning model achieved an AUC of 0.799 ± 0.002 in classifying biopsy-confirmed lesions as being malignant or benign. It can help to further reduce over 23% of unnecessary biopsies while missing only 2% of cancer as the second reader on regions that radiologists have low confidence on. Compared with works performing breast-level classification [12, 15–18, 30, 42], our model can provide prediction for each individual suspicious lesion, and therefore present precise guidance for follow-up procedure including biopsy and surgery.

We acknowledge some limitations of this work. For instance, we did not capture the levels of difficulty related to different types of cancer, which is clinically valuable. We leave this for future work. In addition, the context network we considered in the study did not perform cross-view reasoning, and we expect that networks utilizing all four standard views in a mammogram examination can introduce more complete information and result in more reliable cancer detection.

Conclusion

Besides performing breast-level classification, deep learning methods can help further reduce unnecessary biopsies by classifying suspicious small regions as being benign or malignant. Furthermore, incorporating global image context can improve the network's ability to distinguish between localized benign and malignant lesions on high-resolution images. Future research on techniques for combining local information with global context may be promising for breast cancer screening.

References

1. American Cancer Society: Cancer Facts and Figures. Atlanta, Ga: American Cancer Society, 2020.
2. US Preventive Services Task Force: Medication Use to Reduce Risk of Breast Cancer: US Preventive Services Task Force Recommendation Statement. *JAMA* (2019), 322:857–867.
3. Lee, C. S., Monticciolo, D. L., and Moy, L. Screening guidelines update for average-risk and high-risk women. *AJR Am J Roentgenol* (2020), 214:316–323.

4. Monticciolo, D. L., Newell, M. S., Hendrick, R. E., Helvie, M. A., Moy, L., Monsees, B., Kopans, D., Eby, P. R., and Sickles, E. A. Breast cancer screening for average-risk women: Recommendations from the acr commission on breast imaging. *J Am Coll Radiol* (2017), 14:1137–1143.

5. Oeffinger, K. C., Fontham, E. T. H., Etzioni, R., Herzig, A., Michaelson, J. S., Shih, Y.-C. T., Walter, L. C., Church, T. R., Flowers, C. R., LaMonte, S. J., Wolf, A. M. D., DeSantis, C., Lortet-Tieulent, J., Andrews, K., Manassaram-Baptiste, D., Saslow, D., Smith, R. A., Brawley, O. W., and Wender, R. Breast Cancer Screening for Women at Average Risk: 2015 Guideline Update From the American Cancer Society. *JAMA* (2015), 314:1599–1614.

6. Lehman, C. D., Arao, R. F., Sprague, B. L., Lee, J. M., Buist, D. S., Kerlikowske, K., Henderson, L. M., Onega, T., Tosteson, A. N., Rauscher, G. H., et al. National performance benchmarks for modern screening digital mammography: update from the breast cancer surveillance consortium. *Radiology* (2017), 283:49–58.

7. Ong, M.-S., and Mandl, K. D. National expenditure for false-positive mammograms and breast cancer overdiagnoses estimated at \$4 billion a year. *Health affairs* (2015), 34:576–583.

8. Vlahiotis, A., Griffin, B., Stavros, A. T., and Margolis, J. Analysis of utilization patterns and associated costs of the breast imaging and diagnostic procedures after screening mammography. *ClinicoEconomics and outcomes research: CEOR* (2018), 10:157.

9. Chubak, J., Boudreau, D. M., Fishman, P. A., and Elmore, J. G. Cost of breast-related care in the year following false positive screening mammograms. *Medical care* (2010), 48:815.

10. Fenton, J. J., Taplin, S. H., Carney, P. A., Abraham, L., Sickles, E. A., D'Orsi, C., Berns, E. A., Cutter, G., Hendrick, R. E., Barlow, W. E., et al. Influence of computer-aided detection on performance of screening mammography. *N Engl J Med* (2007), 356:1399–1409.

11. Lehman, C. D., Wellman, R. D., Buist, D. S., Kerlikowske, K., Tosteson, A. N., and Miglioretti, D. L. Diagnostic accuracy of digital screening mammography with and without computer-aided detection. *JAMA Intern Med* (2015), 293:1828–1837.

12. Aboutalib, S. S., Mohamed, A. A., Berg, W. A., Zuley, M. L., Sumkin, J. H., and Wu, S. Deep learning to distinguish recalled but benign mammography images in breast cancer screening. *Clin Cancer Res* (2018), 24:5902–5909.

13. Kim, E.-K., Kim, H.-E., Han, K., Kang, B. J., Sohn, Y.-M., Woo, O. H., and Lee, C. W. Applying data-driven imaging biomarker in mammography for breast cancer screening: preliminary study. *Sci Rep* (2018), 8:1–8.

14. Kyono, T., Gilbert, F. J., and van der Schaar, M. Mammo: A deep learning solution for facilitating radiologist-machine collaboration in breast cancer diagnosis. *arXiv:1811.02661* (2018).

15. McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G. C., Darzi, A., et al. International evaluation of an ai system for breast cancer screening. *Nature* (2020), 577:89–94.

16. Shen, Y., Wu, N., Phang, J., Park, J., Liu, K., Tyagi, S., Heacock, L., Kim, S., Moy, L., Cho, K., et al. An interpretable classifier for high-resolution breast cancer screening images utilizing weakly supervised localization. *arXiv:2002.07613* (2020).

17. Wu, N., Phang, J., Park, J., Shen, Y., Huang, Z., Zorin, M., Jastrzebski, S., Févry, T., Katsnelson, J., Kim, E., et al. Deep neural networks improve radiologists' performance in breast cancer screening. *IEEE Trans Med Imaging* (2019), 39:1184–1194.

18. Zhu, W., Lou, Q., Vang, Y. S., and Xie, X. Deep multi-instance networks with sparse label assignment for whole mammogram classification. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention* (2017), pp. 603–611.

19. Cohen, E. O., Tso, H. H., and Leung, J. W. Multiple bilateral circumscribed breast masses detected at imaging: Review of

evidence for management recommendations. *AJR Am J Roentgenol* (2020), 214:276–281.

20. Leung, J. W., and Sickles, E. A. Multiple bilateral masses detected on screening mammography: assessment of need for recall imaging. *AJR Am J Roentgenol* (2000), 175:23–29.

21. Xi, P., Shu, C., and Goubran, R. Abnormality detection in mammography using deep convolutional neural networks. In *Proceedings IEEE International Symposium on Medical Measurements and Applications* (2018), pp. 1–6.

22. Agarwal, R., Diaz, O., Lladó, X., Yap, M. H., and Martí, R. Automatic mass detection in mammograms using deep convolutional neural networks. *J Med Imaging* (2019), 6:031409.

23. Liu, Y., Zhang, F., Zhang, Q., Wang, S., Wang, Y., and Yu, Y. Cross-view correspondence reasoning based on bipartite graph convolutional network for mammogram mass detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 3812–3822.

24. Ribli, D., Horváth, A., Unger, Z., Pollner, P., and Csabai, I. Detecting and classifying lesions in mammograms with deep learning. *Sci Rep* (2018), 8:1–7.

25. Samala, R. K., Chan, H.-P., Hadjiiski, L., Helvie, M. A., Wei, J., and Cha, K. Mass detection in digital breast tomosynthesis: Deep convolutional neural network with transfer learning from mammography. *Med physics* (2016), 43:6654–6666.

26. He, K., Gkioxari, G., Dollár, P., and Girshick, R. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision* (2017), pp. 2961–2969.

27. Pereira, S. M. P., McCormack, V. A., Moss, S. M., and dos Santos Silva, I. The spatial distribution of radiodense breast tissue: a longitudinal study. *Breast Cancer Res* (2009), 11:R33.

28. Wei, J., Chan, H.-P., Wu, Y.-T., Zhou, C., Helvie, M. A., Tsodikov, A., Hadjiiski, L. M., and Sahiner, B. Association of computerized mammographic parenchymal pattern measure with breast cancer risk: a pilot case-control study. *Radiology* (2011), 260:42–49.

29. Conant, E. F., Barlow, W. E., Herschorn, S. D., Weaver, D. L., Beaber, E. F., Tosteson, A. N. A., Haas, J. S., Lowry, K. P., Stout, N. K., Trentham-Dietz, A., diFlorio Alexander, R. M., Li, C. I., Schnall, M. D., Onega, T., Sprague, B. L., and for the Population-based Research Optimizing Screening Through Personalized Regimen (PROSPR) Consortium. Association of Digital Breast Tomosynthesis vs Digital Mammography With Cancer Detection and Recall Rates by Age and Breast Density. *JAMA Oncol* (2019), 5:635–642.

30. Shen, Y., Wu, N., Phang, J., Park, J., Kim, G., Moy, L., Cho, K., and Geras, K. J. Globally-aware multiple instance classifier for breast cancer screening. In *Proceedings of International Workshop on Machine Learning in Medical Imaging* (2019), pp. 18–26.

31. Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017), pp. 4700–4708.

32. Kingma, D. P., and Ba, J. Adam: A method for stochastic optimization. *arXiv:1412.6980* (2014).

33. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. Imagenet large scale visual recognition challenge. *International journal of computer vision* (2015), 115:211–252.

34. Bergstra, J., and Bengio, Y. Random search for hyper-parameter optimization. *Proc Mach Learn Res* (2012), 13:281–305.

35. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. Pytorch: An imperative style, high-performance deep learning library. In *Proceedings of the Neural Information Processing Systems Conference* (2019), pp. 8026–8037.

36. He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2016), pp. 770–778.

37. Guan, Q., and Huang, Y. Multi-label chest X-ray image classification via category-wise residual attention learning. *Pattern Recogn. Lett. 130* (2020), 130:259–266.

38. Hannun, A. Y., Rajpurkar, P., Haghpanahi, M., Tison, G. H., Bourn, C., Turakhia, M. P., and Ng, A. Y. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nat Med* (2019), 25:65.

39. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., and Summers, R. M. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017), pp. 2097–2106.

40. He, K., Zhang, X., Ren, S., and Sun, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE International Conference on Computer Vision* (2015), pp. 1026–1034.

41. Spanhol, F. A., Oliveira, L. S., Petitjean, C., and Heutte, L. Breast cancer histopathological image classification using convolutional neural networks. In *Proceedings of the International Joint Conference on Neural Networks* (2016), pp. 2560–2567.

42. Elter, M., Schulz-Wendtland, R., and Wittenberg, T. The prediction of breast cancer biopsy outcomes using two cad approaches that both emphasize an intelligible decision process. *J Med Phys* (2007), 34:4164–4172.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.