Deep Neural Networks Improve Radiologists' Performance in Breast Cancer Screening

Nan Wu [1] **Jason Phang** [1] **Jungkyu Park** [1] **Yiqiu Shen** [1] **Zhe Huang** [1] **Masha Zorin** [2] **Stanisław Jastrzębski** [3] **Thibault Févry** [1] **Joe Katsnelson** [4] **Eric Kim** [4] **Stacey Wolfson** [4] **Ujas Parikh** [4] **Sushma Gaddam** [4] **Leng Leng Young Lin** [4] **Kara Ho** [4] **Joshua D. Weinstein** [4] **Beatriu Reig** [4] **Yiming Gao** [4] **Hildegard Toth** [4] **Kristine Pysarenko** [4] **Alana Lewin** [4] **Jiyon Lee** [4] **Krystal Airola** [4] **Eralda Mema** [4] **Stephanie Chung** [4] **Esther Hwang** [4] **Naziya Samreen** [4] **S. Gene Kim** [4] **Laura Heacock** [4] **Linda Moy** [4] **Kyunghyun Cho** [1] **Krzysztof J. Geras** [4 1]

Abstract

We present a deep CNN for breast cancer screening exam classification, trained and evaluated on over 200,000 exams (over 1,000,000 images). Our model achieves an AUC of 0.895 in predicting the presence of cancer in the breast. We attribute the high accuracy of our model to a two-stage training procedure that allows us to use a very high-capacity patch-level network to learn from pixel-level labels alongside a network learning from breast-level labels. Through a study involving 14 readers, we show that our model is as accurate as an experienced radiologist, and that it can improve the accuracy of radiologists' diagnoses when used as a second reader. We further conduct a thorough analysis of our model's performance on different subpopulations of the screening population, model design, training procedure, errors, and properties of its internal representations.

1. Introduction

Breast cancer is the second leading cancer-related cause of death among women in the US. In 2014, over 39 million screening and diagnostic mammography exams were performed in the US. Although mammography is the only imaging test that has been shown to reduced breast cancer mortality (Duffy et al., 2002), there has been discussion regarding the potential harms of screening, including false positive recalls and associated false positive biopsies. Multicenter studies have shown that traditional computer-aided detection in mammography programs do not improve their diagnostic performance (Lehman et al., 2015). Recent developments in deep learning (LeCun et al., 2015) open possibilities for creating a new generation of tools.

This paper makes several contributions. Primarily, we train and evaluate a set of strong neural networks on a mammography dataset with biopsy-proven labels, that is of a massive size by the standards of medical image analysis. We use two complimentary types of labels: breast-level labels indicating whether there is a benign or malignant finding in each breast, and pixel-level labels indicating the location of the findings. To quantify the value of pixel-level labels, we compare a model using only breast-level labels against a model using both breast-level and pixel-level labels. Our best model achieves an AUC of 0.895 in identifying malignant cases and 0.756 in identifying benign cases on the test set reflecting the screening population. In a reader study, we compared the performance of our best model to that of radiologists and found our model to be as accurate as radiologists in terms of AUC. We also found that a hybrid model, taking the average of the probabilities of malignancy predicted by a radiologist and by our neural network, yields more accurate predictions than either of the two separately. This suggests that our model and radiologists learned different aspects of the task and that our model could be effective as a second reader. Finally, we have published the code and weights of our best models online.

2. Data

Our dataset[1] includes 229,426 digital screening mammography exams (1,001,093 images) from 141,473 patients. Each exam contains at least four images, corresponding to the four standard views used in screening mammography.

[1]Center for Data Science, New York University [2]Department of Computer Science and Technology, University of Cambridge [3]Faculty of Mathematics and Information Technologies, Jagiellonian University [4]Department of Radiology, New York University School of Medicine. Correspondence to: Krzysztof J. Geras <k.j.geras@nyu.edu>.

Appearing at the International Conference on Machine Learning AI for Social Good Workshop, Long Beach, United States, 2019.

[1]The dataset is not currently available publicly, however a detailed description of how it was extracted can be found in a technical report (Wu et al., 2019).

Each exam was assigned labels indicating whether each breast was found to have biopsy-proven malignant or benign findings. We have 5,832 exams with at least one biopsy performed within 120 days of the mammogram. Among these, biopsies confirmed malignant findings for 985 (8.4%) breasts and benign findings for 5,556 (47.6%) breasts. 234 (2.0%) breasts had both malignant and benign findings. For the remaining screening exams that were not matched with a biopsy, we assigned labels corresponding to the absence of malignant and benign findings in both breasts.

For all exams matched with biopsies, we asked a group of radiologists to retrospectively indicate the location of the biopsied lesions at a pixel level. An example of such a segmentation is shown in Figure 1. According to the radiologists, approximately 32.8% of exams were mammographically occult, i.e., the lesions that were biopsied were not visible on mammography, even retrospectively, and were identified using other imaging modalities.

3. Deep CNNs for cancer classification

Our goal is to produce predictions corresponding to the four labels for each exam. As input, we take four high-resolution images corresponding to the four standard screening mammography views. See Figure 2 for a schematic overview.

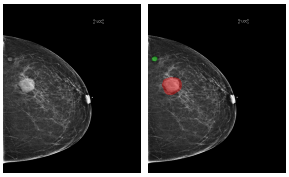

Figure 1. An example of a segmentation. Left: the original image. Right: the image with lesions requiring a biopsy highlighted. The malignant finding is highlighted with red and benign finding with green.

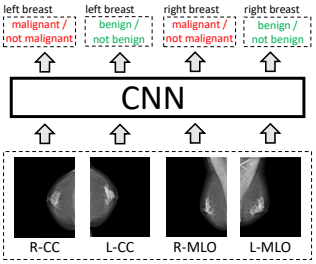

Figure 2. A schematic representation of how we formulated breast cancer exam classification as a learning task.

3.1. Model architecture

We trained a deep multi-view CNN of architecture shown in Figure 3, inspired by Geras et al. (2017). We use an input resolution of 2677×1942 pixels for CC views, and 2974×1748 pixels for MLO views, based on the optimal window size selection procedure Wu et al. (2019). The network consists of two core modules: (i) four view-specific columns, each outputting a fixed-dimension hidden representation for each mammography view, and (ii) two fully connected layers to map from the computed hidden representations to the output predictions. We used four ResNet-22 columns to compute a 256-dimension hidden representation vector of each view. It refers to a 22-layer residual network (He et al., 2016) with additional modifications such as a larger kernel in the first convolutional layer and fewer filters in each layer.

We concatenate the L-CC and R-CC representations into a 512-dimension vector, and apply two fully connected layers to generate predictions for the four outputs. We do the same for the L-MLO and R-MLO views. We average the probabilities predicted by the CC and MLO branches of the model to obtain our final predictions.

3.2. Patch-level classification model and heatmaps

We trained an auxiliary model to classify 256×256-pixel patches of mammograms, predicting the presence or absence of malignant and benign findings in a given patch. The labels for these patches are produced based on overlap with the pixel-level segmentations. We refer to this model as a *patch-level* model, in contrast to the *breast-level* model described above which operates on images of the whole breast. Subsequently, we apply this auxiliary network to the full resolution mammograms in a sliding window fashion to create two 'heatmaps' for each image (Figure 4), containing the estimated probability of malignant and benign findings within a corresponding patch. These heatmaps can be used as additional input channels to the breast-level model to provide supplementary fine-grained information.

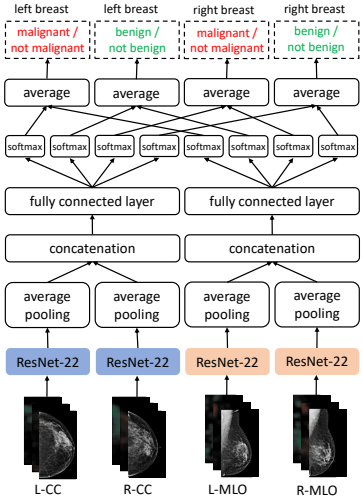

Figure 3. Architecture of our model. The architecture is divided into CC and MLO branches. In each branch, the corresponding left and right representations from the ResNets are individually average-pooled spatially and concatenated, and two fully connected layers are applied to compute the predictions for the four outputs. Weights are shared between L-CC/R-CC columns and L-MLO/R-MLO columns. When heatmaps are added as additional channels to corresponding inputs, the first layers of the columns are modified accordingly.

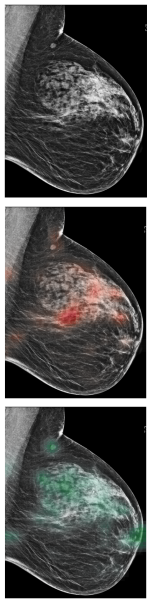

Figure 4. The original image, the 'malignant' heatmap over the image and the 'benign' heatmap over the image.

Using separate breast- and pixel-level models as described above differentiates our work from approaches which utilize pixel-level labels in a single differentiable network (Lot-

ter et al., 2017) or models based on the variations of R-CNN (Ribli et al., 2018). Our approach allows us to use a very deep auxiliary network—a DenseNet121 (Huang et al., 2017)—at the patch level, initialized from pretraining on large off-domain data sets such as ImageNet (Deng et al., 2009), as this network does not have to process the entire high-resolution image at once. Adding the heatmaps produced by the patch-level classifier as additional input channels allows the main classifier to get the benefit from pixel-level labels, while the heavy computation necessary to produce the pixel-level predictions does not need to be repeated each time an example is used for learning. Hereafter, we refer to the model using only breast-level labels as the *image-only* model, and the model using breast-level labels and the heatmaps as the *image-and-heatmaps* model.

4. Experiments

In all experiments, we used the training set for optimizing parameters of our model and the validation set for tuning hyperparameters of the model and the training procedure. We evaluated models with AUC for malignant/not malignant and benign/not benign classification tasks on breast level.

To further improve our results, we applied model ensembling (Dietterich, 2000), wherein we trained five copies of each model with different random initializations of the weights in the fully connected layers. The remaining weights are initialized with the weights of the model pretrained on BI-RADS classification, giving our model a significant boost in performance. For each model, we report the results from a single network (mean and standard deviation across five random initializations) and from an ensemble.

We evaluate our model on several populations to test different hypotheses: (i) *screening population*, including all exams from the test set without subsampling; (ii) *biopsied subpopulation*, which is subset of the screening population, only including exams from the screening population containing breasts which underwent a biopsy; (iii) *reader study subpopulation*, which consists of the biopsied subpopulation and a subset of randomly sampled exams from the screening population without any findings.

4.1. Screening population

We present the results on the screening population, which approximates the distribution of patients who undergo routine screening. Results are shown in the first two rows of Table 1. The model ensemble using only mammogram images achieved an AUC of 0.840 for malignant/not malignant classification and an AUC of 0.743 for benign/not benign classification. The image-and-heatmaps model ensemble using both the images and the heatmaps achieved an AUC of 0.895 for malignant/not malignant and 0.756 for benign/not benign classification, outperforming the image-only model on both tasks. The discrepancy in performance of our models between these two tasks can be largely explained by the fact that a larger fraction of benign findings than malignant findings are mammographically-occult (Table 2). Additionally, there can be noise in the benign/not benign labels associated with radiologists' confidence in their diagnoses. For the same exam, one radiologist might discard a finding as obviously not malignant without requesting a biopsy, while another radiologist might ask for a biopsy.

Table 1. AUCs on screening and biopsied populations.

	single		5x ensemble	
	malignant	benign	malignant	benign
screening population				
image-only	0.827±0.008	0.731±0.004	0.840	0.743
image-and-heatmaps	**0.886±0.003**	**0.747±0.002**	**0.895**	**0.756**
biopsied population				
image-only	0.781±0.006	0.673±0.003	0.791	0.682
image-and-heatmaps	**0.843±0.004**	**0.690±0.002**	**0.850**	**0.696**

We find that the image-and-heatmaps model performs better than the image-only model. Moreover, the image-and-heatmaps model improves more strongly in malignant/not malignant classification than benign/not benign classification. We also find that ensembling is beneficial across all models, leading to a small but consistent increase in AUC.

Table 2. Number of breasts with malignant and benign findings based on the labels extracted from the pathology reports, broken down according to whether the findings were visible or occult.

	malignant		**benign**	
	visible	**occult**	**visible**	**occult**
training	750	107	2,586	2,004
validation	51	15	357	253
test	54	8	215	141
overall	855 (86.8%)	130 (13.2%)	3,158 (56.84%)	2,398 (43.16%)

4.2. Biopsied subpopulation

Results of our models evaluated only on the biopsied subpopulation are in the last two rows of Table 1. Within our test set, this corresponds to 401 breasts: 339 with benign findings, 45 with malignant findings, and 17 with both. This subpopulation that underwent biopsy differs markedly from the overall screening population, which consists of largely healthy individuals undergoing routine annual screening without recall for additional imaging or biopsy.

On the biopsied subpopulation, we observed a consistent difference between the performance of image-only and image-and-heatmaps models. The ensemble of image-and-heatmaps models performs best on both malignant/not malignant classification, attaining an AUC of 0.850, and on benign/not benign classification, attaining an AUC of 0.696. The markedly lower AUCs attained for the biopsied subpopulation, in comparison to the screening population, can be explained by the fact that exams that require a recall for

diagnostic imaging and that subsequently need a biopsy are more challenging for both radiologists and our model.[2]

5. Reader study

To compare the performance of our image-and-heatmaps ensemble (hereafter referred to as *the model*) to human radiologists, we performed a reader study with 14 readers, with varying levels of experience, each reading 720 exams from the test set and providing a probability estimate of malignancy on a 0%-100% scale for each breast in an exam. Among the 1,440 breasts in 720 exams, there are 62 breasts labeled as malignant and 356 breasts labeled as benign. Exams were shuffled before being given to the readers.

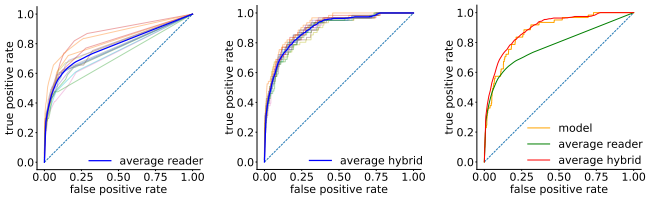

Figure 5. ROC curves for reader study. (left): curves for all 14 readers. Their average performance are highlighted in blue. (middle): curves for hybrid of the image-and-heatmaps ensemble with each single reader. Curve highlighted in blue indicates the average performance of all hybrids. (right): comparison among the image-and-heatmaps ensemble, average reader and average hybrid.

Our model achieved an AUC of 0.876. AUCs achieved by individual readers varied from 0.705 to 0.860 (mean: 0.778, std: 0.0435). Individual ROCs, along with their averages are shown in Figure 5(left). We also evaluated the accuracy of a human-machine hybrid, whose predictions are the averaged predictions of a radiologist and of the model. Hybrids between each reader and the model achieved an average AUC of 0.891 (std: 0.0109) (cf. Figure 5(middle)). These results suggest our model can be used as a tool to assist radiologists in reading breast cancer screening exams and that it captured different aspects of the task compared to experienced breast radiologists.

Additionally, we examined how the network represents the exams internally by visualizing the hidden representations learned by the best image-and-heatmaps model. We visualize two sets of activations: concatenated activations from the last layer of each of the four image-specific columns, and concatenated activations from the first fully connected layer in both CC and MLO model branches. We embed them into a two-dimensional space using UMAP (McInnes et al., 2018) with the Euclidean distance.

[2]More precisely, this difference in AUC can be explained by the fact that while adding or subtracting negative examples to the test population does not change the true positive rate, it alters the false positive rate. False positive rate is computed as a ratio of false positive and negative. Therefore, when adding easy negative examples to the test set, the number of false positives will be growing slower than the number of all negatives, which will lead to an increase in AUC. On the other hand, removing easy negative examples will have a reverse effect and the AUC will be lower.

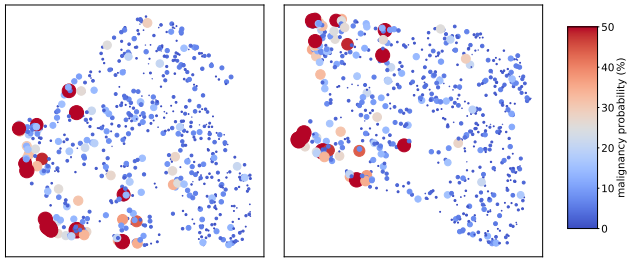

Figure 6. Exams in the reader study set represented using the concatenated activations from the four image-specific columns (left) and the concatenated activations from the first fully connected layer in both CC and MLO model branches (right).

Figure 6 shows the embedded points. Color and size of each point reflect the same information: the warmer and larger the point is, the higher the readers' mean prediction of malignancy is. A score for each exam is computed as an average over predictions for the two breasts. We observe that exams classified as more likely to be malignant according to the readers are close to each other for both sets of activations. The fact that previously unseen exams with malignancies were found by the network to be similar further corroborates that our model exhibits strong generalization capabilities.

6. Discussion

By leveraging a large data set with breast-level and pixel-level labels, we built a neural network which can accurately classify breast cancer screening exams. We attribute this success in large part to the significant amount of computation encapsulated in the patch-level model, which was densely applied to the input images to form heatmaps as additional input channels to a breast-level model. It would be impractical to train this model in a completely end-to-end fashion with currently available hardware. Although our results are promising, we acknowledge that the test set used in our experiments is relatively small and our results require further clinical validation. We also acknowledge that although our model's performance is stronger than that of the radiologists' on the specific task in our reader study, this is not exactly the task that radiologists perform. However, in our study a hybrid model including both a neural network and expert radiologists outperformed either individually, suggesting the use of such a model could improve radiologist sensitivity for breast cancer detection.

In addition, the design of our model is relatively simple. More sophisticated and accurate models are possible. Furthermore, to test the utility of this model in real-time reading of screening mammograms, a clear next step would be predicting the development of breast cancer in the future–before it is even visible to a trained human eye.

References

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, 2009.

Dietterich, T. G. Ensemble methods in machine learning. In *Multiple classifier systems*, 2000.

Duffy, S. W., Tabar, L., Chen, H. H., Holmqvist, M., Yen, M. F., Abdsalah, S., Epstein, B., Frodis, E., Ljungberg, E., Hedborg-Melander, C., Sundbom, A., Tholin, M., Wiege, M., Akerlund, A., Wu, H. M., Tung, T. S., Chiu, Y. H., Chiu, C. P., Huang, C. C., Smith, R. A., Rosen, M., Stenbeck, M., and Holmberg, L. The impact of organized mammography service screening on breast carcinoma mortality in seven swedish counties. *Cancer*, 95(3), 2002.

Geras, K. J., Wolfson, S., Shen, Y., Wu, N., Kim, S. G., Kim, E., Heacock, L., Parikh, U., Moy, L., and Cho, K. High-resolution breast cancer screening with multi-view deep convolutional neural networks. *arXiv:1703.07047*, 2017.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *CVPR*, 2016.

Huang, G., Liu, Z., van der Maaten, L., and Weinberger, K. Q. Densely connected convolutional networks. In *CVPR*, 2017.

LeCun, Y., Bengio, Y., and Hinton, G. Deep learning. *Nature*, 521(7553), 2015.

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 Internal Medicine*, 175(11), 2015.

Lotter, W., Sorensen, G., and Cox, D. A multi-scale CNN and curriculum learning strategy for mammogram classification. In *DLMIA*, 2017.

McInnes, L., Healy, J., Saul, N., and Großberger, L. UMAP: uniform manifold approximation and projection. *J. Open Source Software*, 3(29), 2018. doi: 10.21105/joss.00861. URL https://doi.org/10.21105/joss.00861.

Ribli, D., Horváth, A., Unger, Z., Pollner, P., and Csabai, I. Detecting and classifying lesions in mammograms with deep learning. *Scientific Reports*, 8, 2018.

Wu, N., Phang, J., Park, J., Shen, Y., Kim, S. G., Heacock, L., Moy, L., Cho, K., and Geras, K. J. The NYU breast cancer screening dataset v1.0. Technical report, 2019. Available at https://cs.nyu.edu/~kgeras/reports/datav1.0.pdf.