Deep Neural Networks Improve Radiologists' Performance in Breast Cancer Screening

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

NYU

Center for Advanced Imaging Innovation and Research

Key points

- Breast cancer is the second leading cancer-related cause of death among women in the US.
- We train and evaluate a set of strong neural networks on a dataset of over 200,000 exams (over 1,000,000 images).
- 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.
- 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.
- A hybrid model, taking the average of the probabilities of malignancy predicted by a radiologist and by our network, yields more accurate predictions than either separately.
- The code and weights of our best models are shared on *https://github.com/nyukat/breast_cancer_classifier*.

The NYU Breast Cancer Screening Dataset

Our dataset includes 229,426 screening mammography exams (1,001,093 images) from 141,473 patients. Each exam has labels indicating whether each breast was found to have biopsy-proven malignant or benign findings. For exams matched with biopsies, we asked radiologists to retrospectively indicate the location of the biopsied lesions at a pixel level.

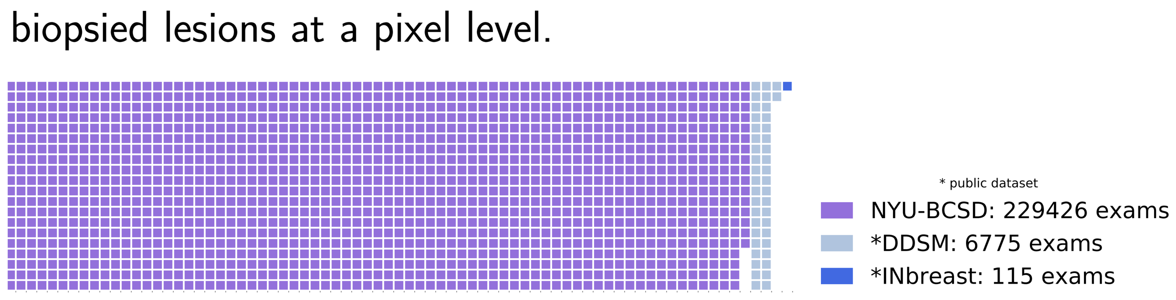

Table 1: Number of breasts with malignant and benign findings based on the labels extracted from the pathology reports, broken down according to whether the findings were visible or occult.

	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%)

Patch-level classifier and heatmaps

We train a network to classify 256×256-pixel patches of mammograms, labeled according to its overlap with the pixel-level segmentation. Subsequently, we apply this network to the full resolution mammograms in a *sliding window fashion* to create two 'heatmaps' for each image, containing the estimated probability of malignant and benign findings within a corresponding patch. Heatmaps can be used as additional input channels to the breast-level classifier.

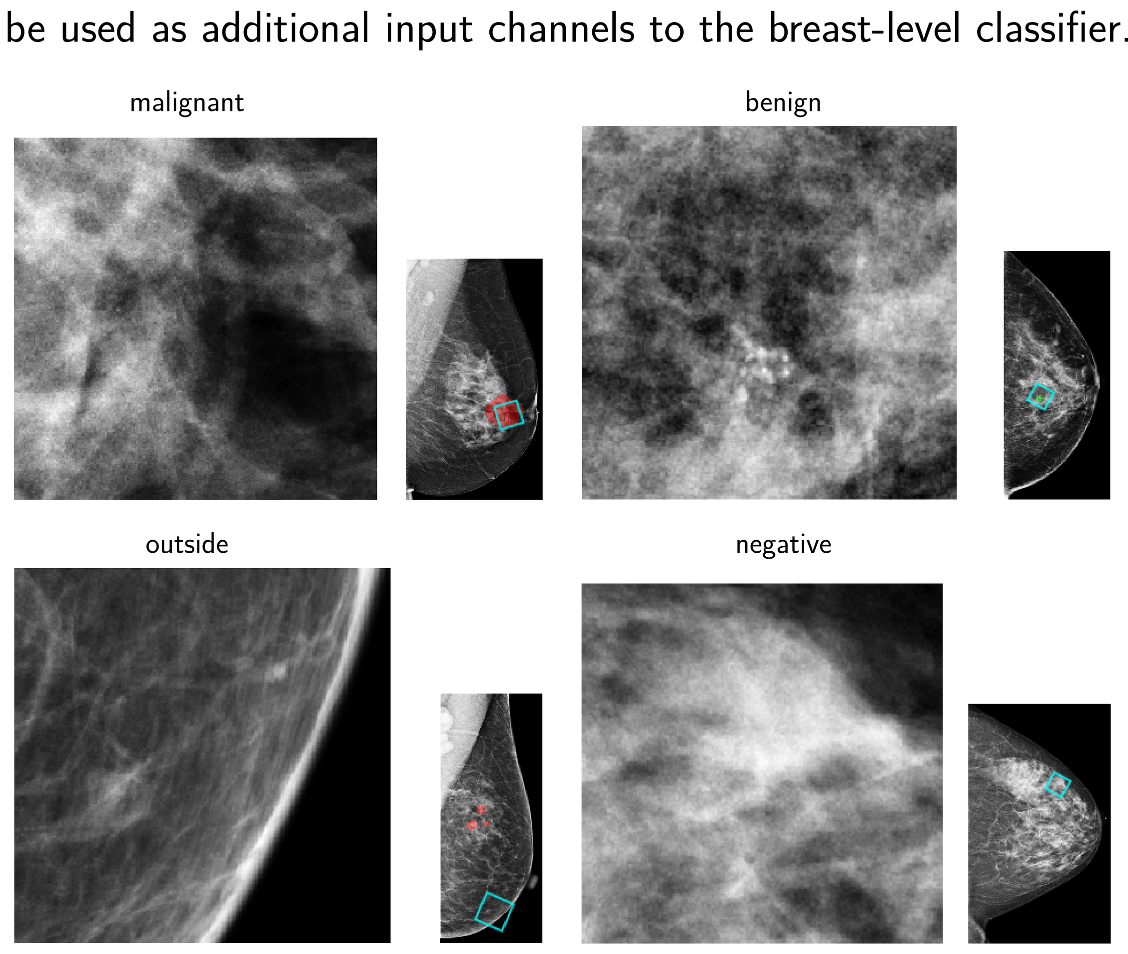

Figure 1: Patches shown with the images they are cropped from. We define four classes: malignant, benign, outside and negative.

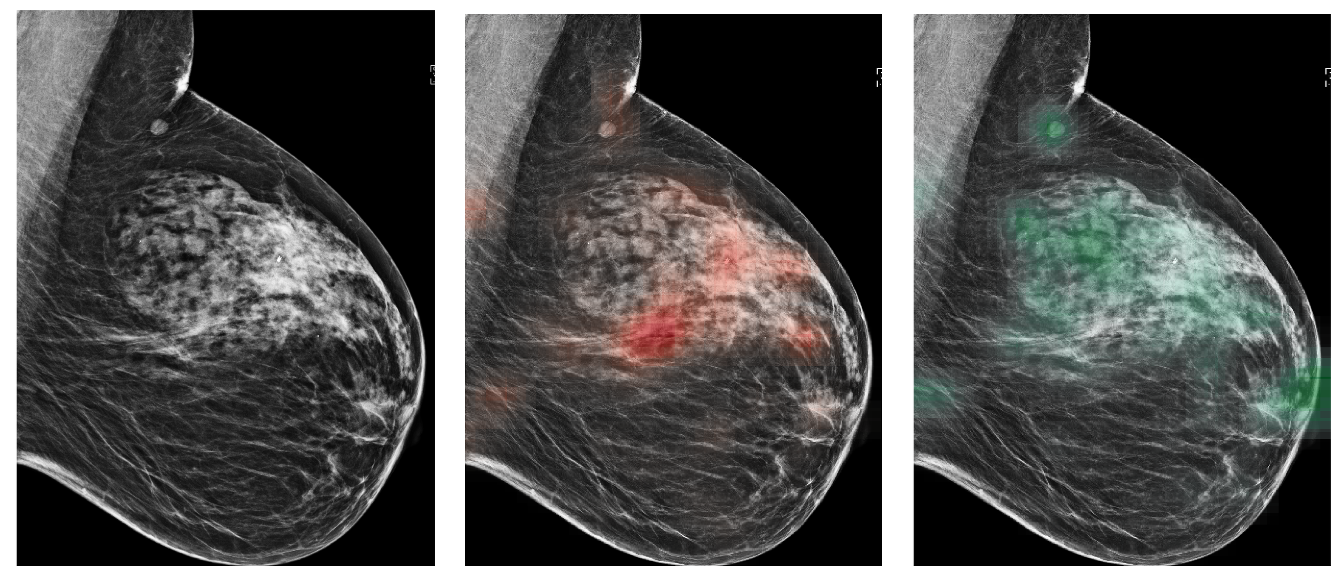

Figure 2: The original image, 'malignant' heatmap and 'benign' heatmap over the image.

Breast-level classifier

We use an input resolution of 2677×1942 pixels for CC views, and 2974×1748 pixels for MLO views. We trained a deep multi-view CNN which consists of two modules: (i) four view-specific columns, each based on the ResNet architecture that outputs a fixed-size representation, and (ii) two fully connected layers to map these representations to predicted probabilities. The ResNet weights are initialized with the weights of the model pretrained on BI-RADS classification [1]. Weights are shared between the L-CC and R-CC columns, as well as the L-MLO and R-MLO columns.

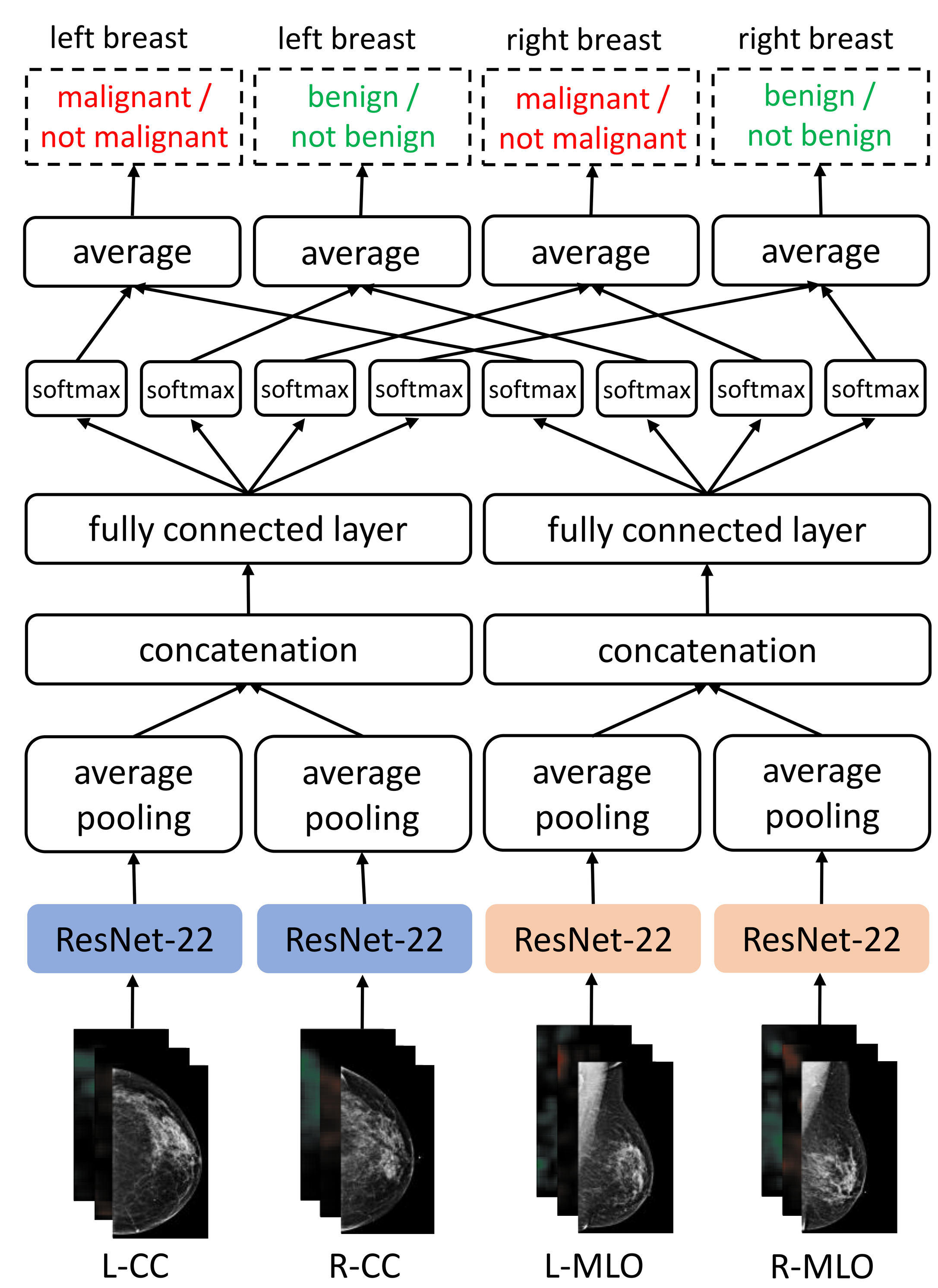

Figure 3: Architecture of our model.

Model evaluation

We evaluate our model on the following populations:

- **screening population**, the entire test set without subsampling;
- **biopsied subpopulation**, a subset of the screening population, only including exams containing breasts which underwent a biopsy;

Table 2: 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**

The markedly lower AUCs attained for the biopsied subpopulation, in comparison to the screening population, can be explained by the fact that exams subsequently requiring a biopsy are more challenging for both radiologists and our model. The heatmaps help more strongly in the malignant/not malignant classification task. This discrepancy can be largely explained by the fact that a larger fraction of benign findings than malignant findings are mammographically-occult (Table 1).

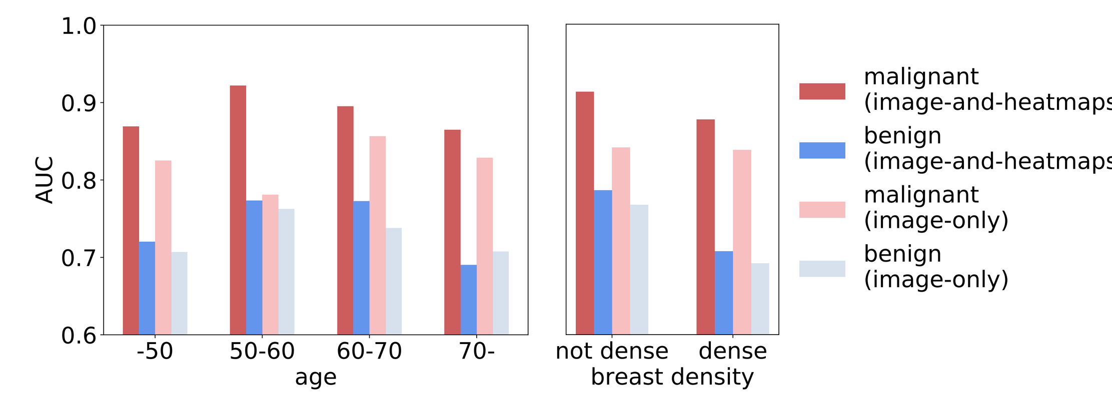

Figure 4: AUCs for patients grouped by age and by breast density.

Comparison to human radiologists

Reader study subpopulation consists of the biopsied subpopulation and equal number of randomly sampled exams from the screening population without any findings. On this subpopulation, we performed a reader study with 14 radiologists, each reading all exams and providing a probability estimate of malignancy on a 0%-100% scale for each breast in an exam.

- 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).
- Human-machine hybrids, whose predictions are the averaged predictions of a radiologist and of the model, achieved an average AUC of **0.891** (std: 0.0109).

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 radiologists.

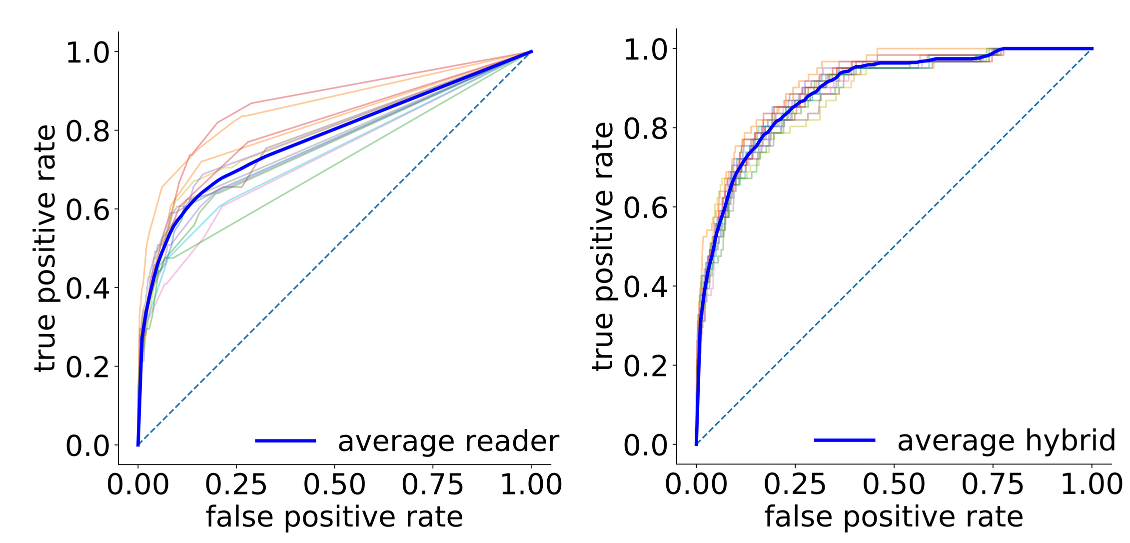

Figure 5: ROC curves for all readers (left). ROC curves for hybrid of the model with each single reader (right). Curve highlighted in blue indicates the average performance.

We examined how the network represents the exams by visualizing the hidden representations learned by the model. 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. The fact that previously unseen exams with malignancies were found by the network to be close in this low-dimensional space further corroborates that our model exhibits strong generalization capabilities.

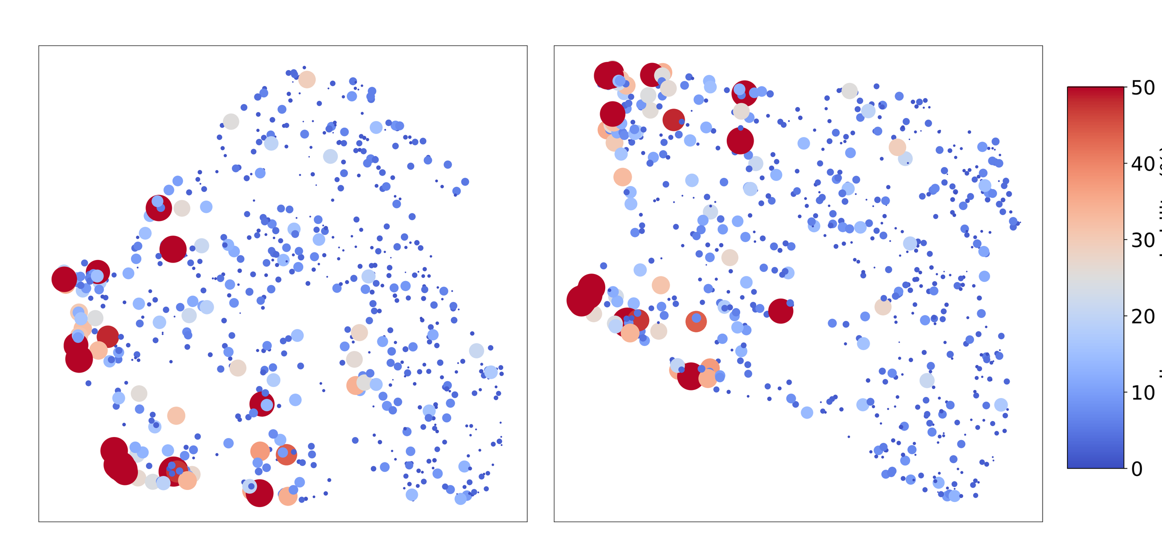

Figure 6: Exams represented using the concatenated activations from the four image-specific columns (left) and the concatenated activations from the first fully connected layer in both CC and MLO model branches (right).

The full paper

This is a shorter version of the paper available at *https://arxiv.org/pdf/1903.08297.pdf*.

References

[1] Krzysztof J. Geras, Stacey Wolfson, Yiqiu Shen, Nan Wu, S. Gene Kim, Eric Kim, Laura Heacock, Ujas Parikh, Linda Moy, and Kyunghyun Cho, "High-resolution breast cancer screening with multi-view deep convolutional neural networks," *arXiv:1703.07047*, 2017.