Augmenting Computer Aided Detection of Malignancies in Mammograms with Radiologist Input

Skylar W Wurster1 Jian Chen1 Karla Evans2 Arkadiusz Sitek3 Jeremy M Wolfe4\*

Machine learning models in the medical field currently underutilize some of the most valuable information available – input from the medical professionals themselves. Radiologists perform better than chance even when given only fractions of a second to classify a mammogram as normal or abnormal. We implement a deep learning model and use radiologist “gist” input to achieve an AUC of 0.87 - higher than the radiologists or the model alone.

Radiologist currently utilize assistance from automatic computer interpretation or Computer Aided Detection (CAD) systems that support radiologists in making decisions [2-5]. These systems have been useful by the side of radiologists, but work has yet to be done to use radiologist input within the CAD system itself. This is a serious detriment to the advancement of CAD systems and breast cancer detection, as skilled professionals’ inputs are not being utilized within those CAD systems. Like the idea behind transfer learning, we utilize the knowledge that the radiologists already have to speed up and improve results.

We hypothesize that by using radiologist opinions within a machine learning model used to predict breast cancer from mammography scans, we will see improved results over just the radiologist or the machine alone. Our approach utilizes radiologist gist readings within a CAD system that uses a deep convolutional neural network (CNN) and transfer learning to detect subtle abnormalities in full-field mammography. Radiologist gist data was taken by exposing radiologists to a unilateral mammogram of a breast with no abnormalities, a breast with an abnormality, or a breast contralateral to a breast with an abnormality for 250 milliseconds [6]. This short exposure time is called gist perception in vision science literature. They were asked to rate the mammogram from 100 to 0, where 100 is normal and 0 is abnormal. We see substantial improvements to the CAD system when implementing a voting system that takes radiologist input into account within the algorithm, and even improvements over the radiologists.

\* jwolfe@bwh.harvard.edu

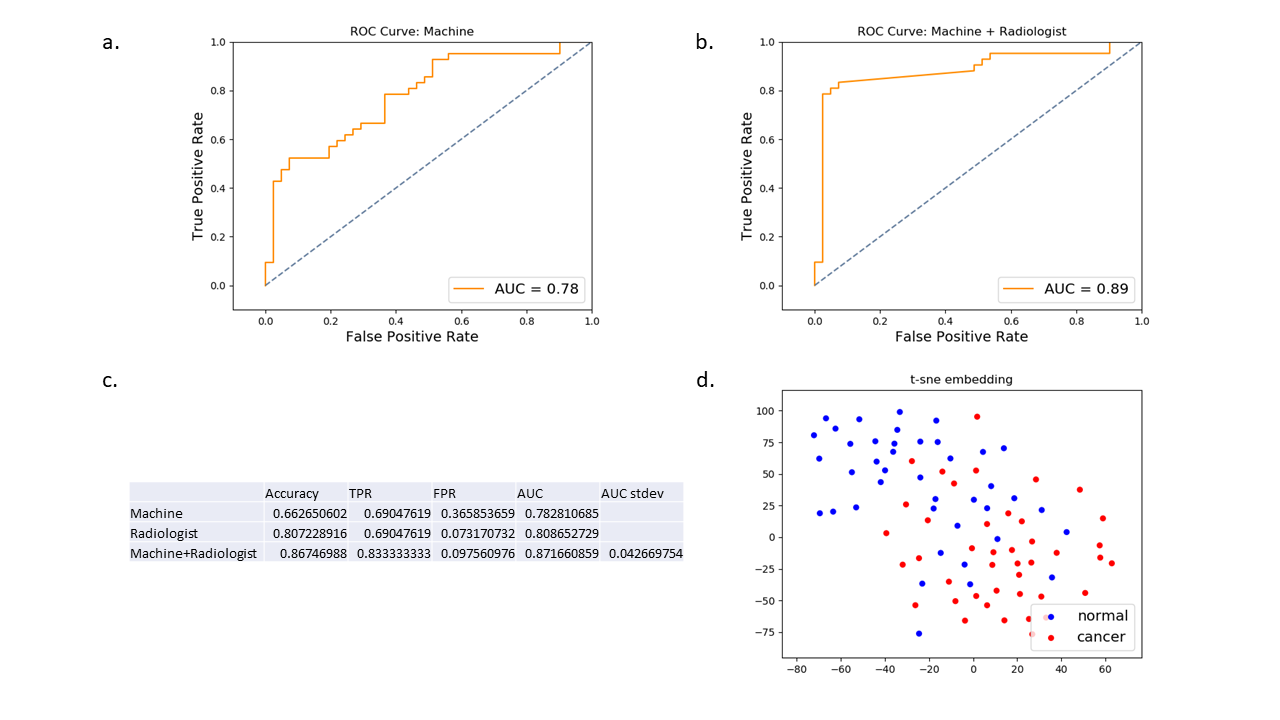
1: Ohio State University

2: University of York

3: IMB Research

4: Harvard University

The radiologists perform well at classifying abnormal and normal mammograms in our dataset, with an accuracy of 80.7% and an area on the receiver operating characteristic (AUC) of 0.808. The classifier by itself performs worse than the radiologists with AUC=0.77. The voting system improves upon both the radiologist and the model’s output, with an AUC=0.88 and an accuracy of 86.7%.



1. **a.** ROC curve for the classifier alone. **b.** ROC curve for the voting system using both the machine and radiologist input. **c.** Accuracy, true positive rate (TPR), false positive rate (FPR), area under the receiver operating characteristic (AUC) and the standard deviation of the AUC for the three parties/systems predicting abnormal mammograms. **d.** t-SNE visualization of the voting system based on the classification, confidence score, and a 50-dimensional vector generated by PCA on the deep visual features of each image.

We notice interesting trends when comparing the radiologist decisions to the voting model’s decisions. The classifier manages to correct 10 of the 16 errors that the radiologists made, while only introducing 2 false positives and 2 false negatives. These corrected errors include 1 false positive and 9 false negatives.

On the other hand, the radiologists corrected 18 errors the classifier made, of which were 7 false negatives and 11 false positives. There are also 5 instances of both the classifier and the radiologists making the same mistake. The radiologists only introduce 1 mistake, a false negative, into the voting system.

On average, radiologists had a confidence of 0.346 and the classifier had a confidence of 0.554, showing the classifier was more confident in results than the radiologists. When the radiologists corrected the machine correctly, they had an average confidence of 0.473. When the classifier corrected the radiologist correctly, it had an average confidence of 0.811. We observe little correlation between the confidence values, with a Pearson correlation coefficient of r=-0.09, and a correlation of r=0.2 between predictions.

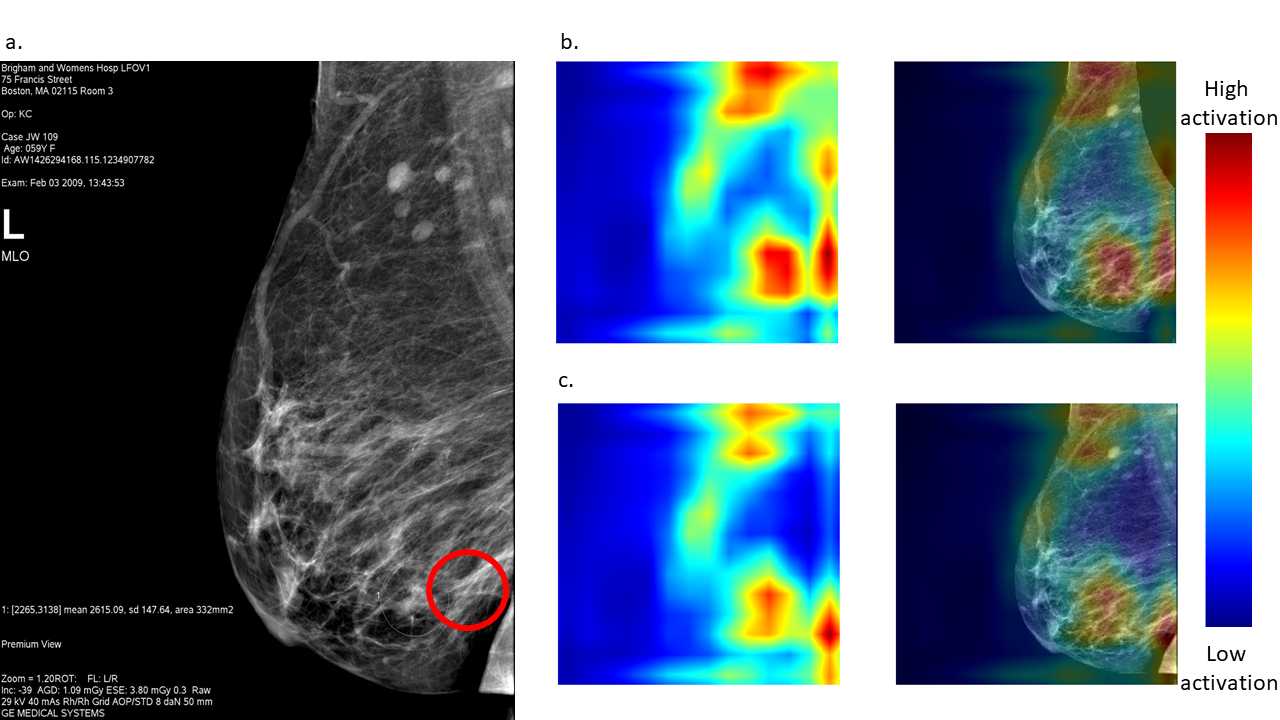


Figure 2: **a.** A radiologist annotation on a mammogram of a breast with a malignancy. The region of interest (ROI) is marked with the red circle around it. **b.** Grad-CAM++ activation map of the model’s decision for the same mammogram as shown in a. that is cropped to remove the pectoral muscle and noise along the bottom. There is a large activation in the ROI. **c.** Grad-CAM++ activation map of the model’s decision for the mammogram left uncropped. We see more activation from the pectoral muscle and noise, and less activation on the ROI than in b.

We see an improvement over a typical classifier when incorporating radiologist input, improving AUC from 0.78 to 0.87. The slight improvement over the radiologist performance shows a strong possibility for machine learning algorithms and human decisions to complement each other to improve overall performance. This claim is supported by the weak correlation between confidence values and predictions for the classifier versus the radiologists – the “gist” extracted from the radiologist must capture different features than the deep learning model captures. Despite the classifier confidence being higher than the radiologists on average, very few incorrect classifications were introduced. This means the classifier was not confident enough with its false predictions allowing the radiologists with more experience to give correct input. There are also several instances in which both the radiologists and the classifier are incorrect with their prediction. These cases may be more challenging to evaluate. This is supported by the fact that the average confidence for these incorrect classifications was 0.19 over both the radiologists and classifier, meaning neither party was confident in their prediction.

Our approach incorporates the informed decisions of radiologists with that have years of education and experience with the image analysis and pattern recognition capabilities provided by deep learning and other machine learning techniques. The combination of the parts is better than either solution alone. Similar solutions utilizing both inputs may prove useful for other problem domains, especially in the medical field where trained professionals work with computer aided detection systems often.

# Methods

The mammography dataset contains 220 unilateral full-field digital mammograms obtained from 110 unique patients at Brigham and Women’s Hospital. Three classes are in these images – no malignancy (110 images), malignancy (66 images), or contralateral to the breast with a malignancy (44 images). The left breast mammograms are mirrored. The images are cropped such that muscle tissue is removed in an algorithm like that suggested in [9] to increase stability on the relatively small dataset. Once cropped, we visually inspect the results to ensure the cropping didn’t remove too much or too little. Errors are corrected by hand. All code for this paper is provided on Github: <https://github.com/skywolf829/DeepMammo>.

The radiologist data is comprised of responses from 10 radiologists who were each shown 120 images of the 220 where 40 were completely normal, 40 had a confirmed malignancy, and 40 were normal but contralateral to a breast with a malignancy. They were shown the image for 250 milliseconds and asked to report on a scale from 0, recommending the patient return for further examination, to 100, the scan is normal. The readings from the 10 radiologists were averaged for each image, giving the final radiologist response.

The final classification for the radiologist is set to 0 (normal) if the average response is greater than 50 and set to 1 (malignant) otherwise. A “confidence score” for each image is calculated based on the average response’s difference from 50, then normalized.

For the CAD system, a transfer learning model based on VGG19 was used [8]. VGG19 is a deep CNN trained to classify over 1000 classes from the ImageNet corpus containing over 10 million non-medical images. Values from nodes containing deep visual features located in the first fully connected layer after the convolutional layers (relu6) are used capture features about a photo fed through the network. Those features are used as inputs into a supervised learning algorithm, in this case a linear support vector machine (C=0.0001), for binary classification as normal or abnormal.

From the dataset, we use only the normal and malignant images for training and testing. Of these images, only those with radiologist responses are kept, leaving 83 images. Images were downscaled to 244x244 pixels to fit the pre-trained VGG19 model properly. Leave one out training and testing is used to generate 83 out-of-fold predictions and confidence values for each image in the set. Predictions will be 0 if classified as normal, and 1 if classified as malignant. Confidence scores are the absolute value of the dot product of the weights learned by the linear SVM and the feature vector for the image.

A simple voting system is used which will take the response that has a higher confidence between the radiologist and transfer learning model. The prediction from the party with the higher confidence is taken as the voting system’s prediction. Bootstrapping 1000 AUC samples on the final predictions is used to estimate the variance of the area under the receiver operating characteristic (AUC).

References

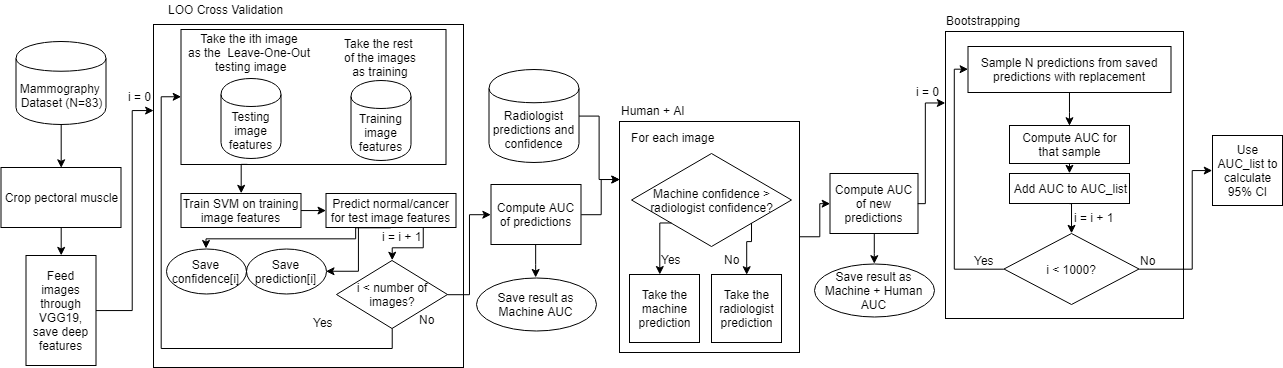
1. M. Kalager, M. Zelen, F. Langmark, and H. O. Adami. New England Journal of Medicine, 363(13):1203–1210, 2010.
2. B. Q. Huynh, H. Li, and M. L. Giger. Journal of Medical Imaging, 3(3):034501–034501, 2016.
3. Giger M. L., Boone J., Chan H.,. Med. Phys. 35(12), 5799–5820 (2008).
4. Hadjiiski L., Sahiner B., Chan H.-P., Curr. Opin. Obstet. Gynecol. 18(1), 64–70 (2006). 10.1097/01.gco.0000192965.29449.da
5. Giger M. L., Karssemeijer N., Schnabel J. A., Annu. Rev. Biomed. Eng. 15, 327–357 (2013).10.1146/annurev-bioeng-071812-152416
6. Evans K. K., Haygood T.M., Cooper J., Culpan A.-M., Wolfe J.M. Proceedings of the National Academy of Sciences of the United States of America. 2016;113(37):10292–10297. doi: 10.1073/pnas.1606187113.K.
7. K. Simonyan and A. Zisserman. CoRR, vol. abs/1409.1556, 2014.
8. A. Rampun, P. J. Morrow, B. W. Scotney, J. Winder. Artificial Intelligence in Medicine, Volume 79, 2017, Pages 28-41, ISSN 0933-3657

# Author Contributions

Original dataset supplied by K.E. Idea and guidance offered by J.M.W and A.S. Advising by J.C. Implementation and testing by S.W.W, whom also wrote the manuscript.

# Competing Interests

The authors declare no competing interests.



Supplementary Figure 1: Diagram of the methods used in this experiment. All code and data available online.