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 can 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 0.91 - higher than the radiologists or the model alone.

The American Cancer Society estimates that there will be over 270,000 new cases of breast cancer in 2019. If these diagnoses are made soon after the cancer first develops through mammography screening, the chance of survival is over 90% [1]. Radiologist have had 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 on utilizing 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 aren’t being utilized within those CAD systems.

We hypothesize that by combining machine learning output with radiologist opinions on mammography scans, we will see improved results when asking for classification of new mammography scans. 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 model.

\* [Wurster.18@osu.edu](mailto:Wurster.18@osu.edu)

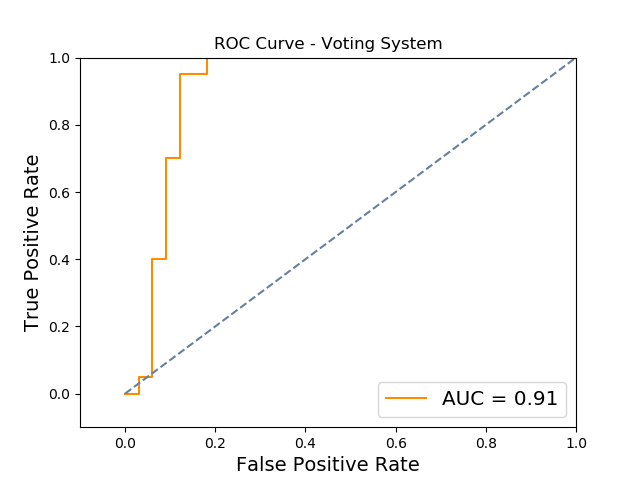
1: Ohio State University

2: University of York

3: IMB Research

4: Harvard University

On the test set, the radiologists perform exceptionally well at classifying abnormal and normal mammograms, with an accuracy of 89% and an area on the receiver operating characteristic (AUC) of 0.87. The classifier by itself performs worse than the radiologists, but still respectably with AUC=0.69. The voting system improves upon both the radiologist and the model’s output, with an AUC=0.91, as shown in Figure 1.



1. The ROC curve for our voting system.

We notice interesting trends when comparing the radiologist decisions to the voting model’s decisions. The voting model manages to correct five of the six errors that the radiologists made, while only introducing four false positives. These corrected errors include two false positives and three false negatives. These corrections were made on images in which the radiologist confidence was lower than the classifier confidence. The radiologist confidence values for these corrected images were very low, each scoring less than 0.2. On average, radiologists had a confidence of 0.358 and the classifier had a confidence of 0.306, showing the classifier was slightly less confident in results than the radiologists. Despite a similar average confidence, there is little correlation between the confidence values, with a Pearson correlation coefficient of r=0.060.

We see a large improvement over a typical classifier when incorporating radiologist input, improving AUC from 0.69 to 0.91. 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 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 lower than the radiologist on average, it was able to identify and correct five of the six errors the radiologists made. The only unresolved error was a false negative, in which both the radiologists and classifier had a very low confidence in their answers, implying that it was a difficult case.

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 images are cropped such that muscle tissue is removed in an algorithm similar to that suggested in [9] to increase stability on the relatively small dataset. All code for this paper is provided on Github: <https://github.com/skywolf829/DeepMammo>.

For the radiologist gist data, 10 radiologists were 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 that are deep in the network are used capture features about a photo that was put into the network. Those features are used as inputs into a supervised learning algorithm, in this case a linear support vector machine, for binary classification as normal or abnormal.

From the dataset, we use only the normal and malignant images for training and testing. Images were downscaled to 244x244 pixels to fit the pre-trained VGG19 model properly. A 70/30 split is used for training and testing, giving a training set of 123 images and a test set of 53 images. 5-fold cross validation on the training set is used for hyperparameter tuning, where we find C=0.0001 performed best. Confidence values are obtained for each image, defined as the output’s distance from the hyperplane separating the classes. Confidence is capped at 1.

A simple voting system is used which will take the response that has a higher confidence between the radiologist and transfer learning model.

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.