Augmenting Computer Aided Screening of Malignancies in Mammograms with Radiologist Gist 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 0.889 - higher than the radiologists or the model alone.

Gist processing is the visual perception skill humans use to quickly assess perceptual and semantic information from a scene through only a glance that lasts a fraction of a second. This processing allows us to intelligently allocate focus and attention where needed within the scene for the task at hand. In the short time it takes to do gist processing, the visual system forms a spacial representation, the category of the scene, and other global structural information. The gist signal is coarse and global, extracted from the entirety of the scene. Studies have shown that when trained radiologists are tasked to use gist processing for screening mammograms, they perform better than chance, even with only 250 milliseconds of observation.

Concurrently, radiologist 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 and perform well, 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 with features from a deep convolutional neural network (CNN) used to train LightGBM to screen mammograms for subtle abnormalities. 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 baseline model when adding the radiologist gist input into the model’s input, with this new system even outperforming radiologists.

\* jwolfe@bwh.harvard.edu

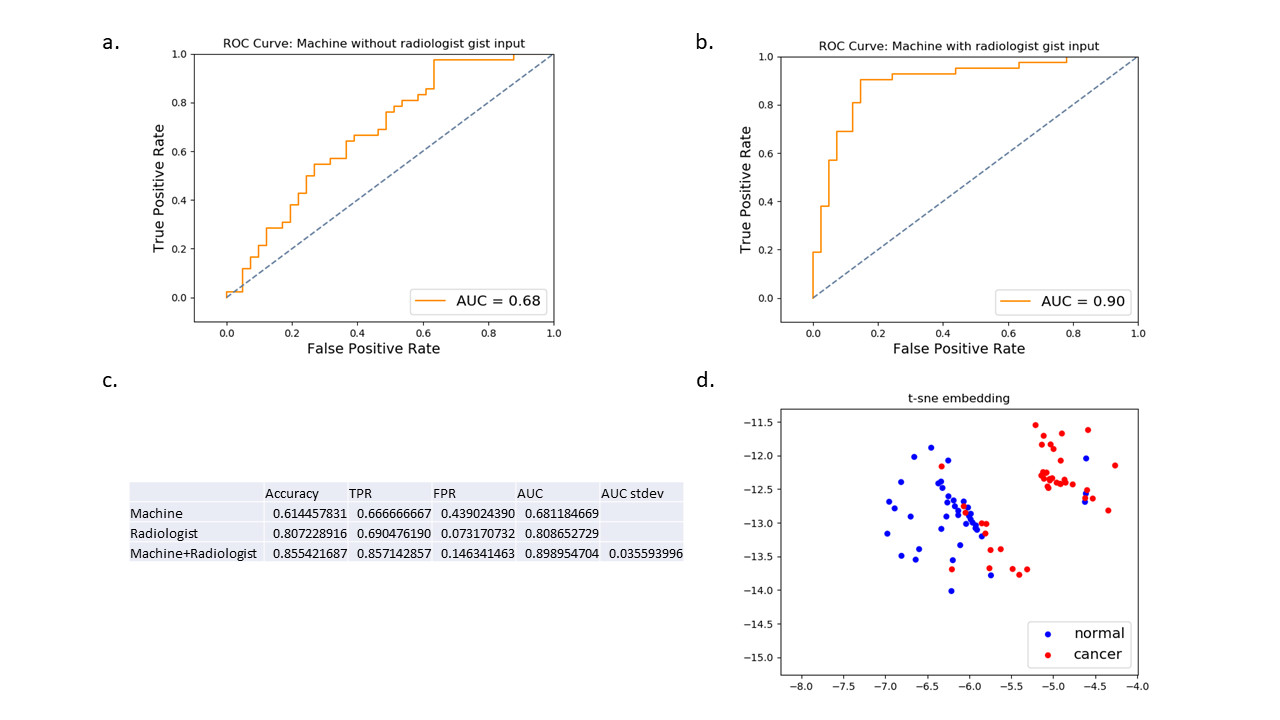
1: Ohio State University

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=0.681. The model incorporating the radiologist gist response improves upon both the radiologist and the model’s output, with an AUC=0.899 and an accuracy of 85.5%.



1. **a.** ROC curve for the classifier without gist input added to the feature vector. **b.** ROC curve for the classifier using both the machine’s deep features with the radiologist gist 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 using the classifier with radiologist gist input.

We notice interesting trends when comparing the radiologist decisions to the classifier’s decisions. The classifier utilizing both the deep features and the radiologist gist input manages to correct 8 of the 16 errors that the radiologists made, while only introducing 3 false positives and 1 false negative. All radiologist errors corrected were false negatives.

On the other hand, after introducing the radiologist gist input to the classifier, 25 errors were corrected, with only 3 errors introduced. These corrections include 8 false negatives and 17 false positives. There are also 5 instances of both the classifier and the radiologists making the same mistake. The radiologists only introduce 3 mistakes - 2 false positives and 1 false negative.

We see an improvement over a typical classifier when incorporating radiologist gist input, improving AUC from 0.681 to 0.899(95% confidence interval 0.823 to 0.962). Before introducing radiologist gist input into the classifier, we see a Pearson correlation coefficient of 0.113 between predictions by the classifier and predictions by the radiologists. After introducing radiologist gist input into the classifier, we see a Pearson correlation coefficient of 0.733. Though these are highly correlated, the significant difference in AUC shows that the deep learned features from Inception V4 must capture additional information about the mammogram that radiologist gist responses aren’t, helping the classifier screen the mammograms with a higher AUC and accuracy than radiologist input alone.

Of all 16 radiologist mistakes, we see an average confidence of 0.178 – lower than the overall average radiologist confidence of 0.346 (p<0.05). Of the 8 mammograms that were incorrectly classified by the radiologist gist response, and then corrected by the classifier, we observe an average confidence of 0.195 which is also lower than the average radiologist confidence (p<0.05). Yet, there are 9 predictions that have a radiologist confidence less than 0.2 in which the classifier is corrected when the radiologist gist input is present. This shows that the classifier is doing more than simply using high-confidence results provided by the radiologist gist response and using the model to classify low-confidence results. Instead, there is a more complex relationship learned between the deep features and radiologist gist input.

Our approach incorporates the informed decisions of radiologists’ gist 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. Though helpful here, we suggest this only be used for screening and not diagnosis. 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. All code and the dataset for this paper are 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. This is experiment 2 from [6]. 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 machine learning model, a transfer learning model based on InceptionV4 was used [8]. InceptionV4 is a deep CNN trained to classify over 1000 classes from the ImageNet corpus containing over 10 million non-medical images. We use the python package pretrainedmodels to load InceptionV4. Values from nodes containing deep visual features located in the first fully connected layer after the convolutional layers are used capture features about a photo fed through the network. Those 1536 features are used as inputs into a supervised learning algorithm, in this case Microsoft’s LightGBM, for binary classification as normal (0) or abnormal (1). LightGBM (LGBM) is a gradient boosting framework using tree-based learning algorithms. We use parameters as follows: learning rate = 0.003, boosting\_type = gbdt, objective = binary, metric = binary\_logloss, sub\_feature = 0.5, num\_leaves = 10, min\_data = 1, max\_depth = 1000. Code is implemented in Python 3.6 with Keras and Tensorflow.

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. This includes 42 abnormal mammograms and 41 normal. Images were downscaled to 299x299 pixels to fit the pre-trained InceptionV4 model properly via a function in pretrainedmodels. 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.

Using InceptionV4 features as inputs into LightGBM gives us our machine results before gist input. Then, for each image’s feature vector of length 1536, we append the radiologist gist classification (0 or 1) and the calculated radiologist gist confidence (0.0-1.0). This creates a new feature vector of length 1538 for each image. These new feature vectors are used as input to LightGBM and is referred to as the machine using gist input. Bootstrapping 1000 AUC samples on the final predictions is used to estimate the variance of the area under the receiver operating characteristic (AUC).

References

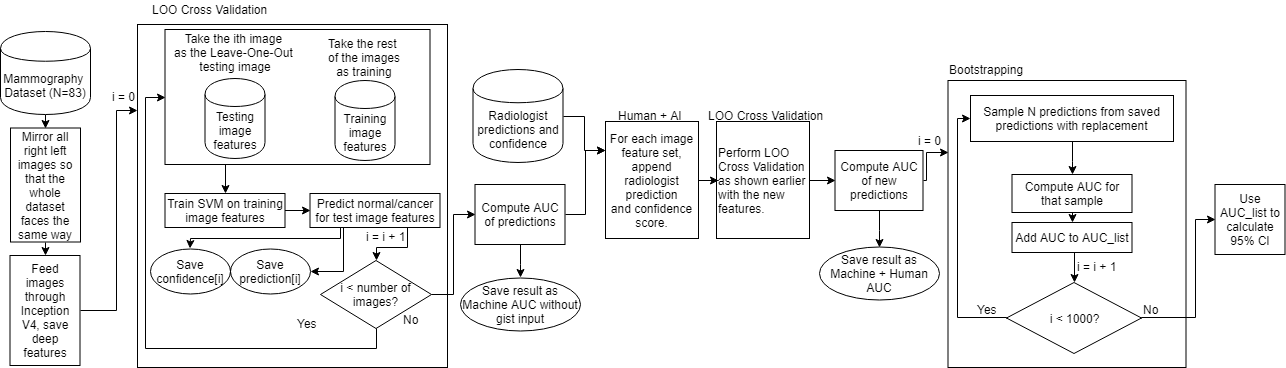
1. M. Kalager, M. Zelen, F. Langmark, and H. O. Adami. New England Journal of Medicine, 363(13):1203–1210, 2010.
2. B. Q. Huynh, H. Li, and M. L. Giger. Journal of Medical Imaging, 3(3):034501–034501, 2016.
3. Giger M. L., Boone J., Chan H.,. Med. Phys. 35(12), 5799–5820 (2008).
4. Hadjiiski L., Sahiner 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
9. Evans K. K., Georgian-Smith D., Tambouret R., Birdwell R. L., Wolfe J. M. Psychon Bull Rev (2013) 20:1170-1175. 10.3758/s13423-013-0459-3
10. Evans K. K., Schill H., Culpan AM., Wolfe J. Journal of Vision 2017;17(10):927. doi: 10.1167/17.10.927.

# 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.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Image | Ground Truth | Machine without gist | Radiologist Prediction | Machine with gist | Radiologist  Confidence  Values | |  |
| N10\_L.png | 0 | 1 | 0 | 0 | 0.462857142 | Same mistake | 5 |
| N10\_R.png | 0 | 1 | 0 | 0 | 0.305714286 | Radiologist corrects model | 25 |
| N11\_L.png | 0 | 0 | 0 | 0 | 0.546666666 | Model corrects radiologist | 8 |
| N11\_R.png | 0 | 1 | 0 | 0 | 0.405714286 | Radiologist introduces error | 3 |
| N12\_L.png | 0 | 1 | 0 | 0 | 0.182857142 | Model introduces error | 4 |
| N12\_R.png | 0 | 0 | 0 | 0 | 0.444 |  |  |
| N13\_L.png | 0 | 0 | 0 | 0 | 0.298571428 |  |  |
| N13\_R.png | 0 | 0 | 0 | 0 | 0.554666666 |  |  |
| N14\_L.png | 0 | 0 | 0 | 0 | 0.326666666 |  |  |
| N14\_R.png | 0 | 0 | 0 | 0 | 0.342857142 |  |  |
| N15\_L.png | 0 | 0 | 1 | 1 | 0.066666666 |  |  |
| N15\_R.png | 0 | 0 | 0 | 0 | 0.247692308 |  |  |
| N16\_L.png | 0 | 1 | 0 | 0 | 0.041428572 |  |  |
| N16\_R.png | 0 | 0 | 0 | 0 | 0.475384616 |  |  |
| N17\_L.png | 0 | 0 | 0 | 0 | 0.549333334 |  |  |
| N17\_R.png | 0 | 0 | 1 | 1 | 0.084 |  |  |
| N18\_L.png | 0 | 0 | 0 | 0 | 0.424 |  |  |
| N18\_R.png | 0 | 0 | 0 | 0 | 0.347142858 |  |  |
| N19\_L.png | 0 | 0 | 0 | 1 | 0.618333334 |  |  |
| N19\_R.png | 0 | 1 | 0 | 1 | 0.341538462 |  |  |
| N1\_L.png | 0 | 1 | 0 | 0 | 0.370666666 |  |  |
| N1\_R.png | 0 | 1 | 0 | 0 | 0.108 |  |  |
| N23\_L.png | 0 | 0 | 0 | 0 | 0.627142858 |  |  |
| N23\_R.png | 0 | 0 | 0 | 0 | 0.553333334 |  |  |
| N2\_L.png | 0 | 1 | 0 | 0 | 0.372307692 |  |  |
| N2\_R.png | 0 | 0 | 0 | 0 | 0.104 |  |  |
| N3\_L.png | 0 | 0 | 0 | 0 | 0.241538462 |  |  |
| N3\_R.png | 0 | 0 | 0 | 1 | 0.317333334 |  |  |
| N4\_L.png | 0 | 1 | 0 | 0 | 0.596 |  |  |
| N4\_R.png | 0 | 1 | 0 | 0 | 0.621428572 |  |  |
| N50\_L.png | 0 | 0 | 0 | 0 | 0.157142858 |  |  |
| N5\_L.png | 0 | 1 | 0 | 0 | 0.344 |  |  |
| N5\_R.png | 0 | 1 | 0 | 0 | 0.545333334 |  |  |
| N6\_L.png | 0 | 1 | 0 | 0 | 0.621428572 |  |  |
| N6\_R.png | 0 | 1 | 0 | 0 | 0.538571428 |  |  |
| N7\_L.png | 0 | 1 | 0 | 0 | 0.493333334 |  |  |
| N7\_R.png | 0 | 1 | 0 | 0 | 0.133333334 |  |  |
| N8\_L.png | 0 | 0 | 0 | 0 | 0.401428572 |  |  |
| N8\_R.png | 0 | 0 | 0 | 0 | 0.557333334 |  |  |
| N9\_L.png | 0 | 1 | 0 | 0 | 0.423076924 |  |  |
| N9\_R.png | 0 | 0 | 1 | 1 | 0.154285714 |  |  |
| AD11\_R.png | 1 | 1 | 1 | 1 | 0.445714286 |  |  |
| AD12\_L.png | 1 | 1 | 0 | 1 | 0.375384616 |  |  |
| AD13\_L.png | 1 | 1 | 1 | 1 | 0.592857142 |  |  |
| AD14\_L.png | 1 | 1 | 0 | 1 | 0.395714286 |  |  |
| AD15\_R.png | 1 | 1 | 1 | 1 | 0.566666666 |  |  |
| AD16\_R.png | 1 | 0 | 0 | 0 | 0.294285714 |  |  |
| AD17\_L.png | 1 | 0 | 1 | 1 | 0.14 |  |  |
| AD18\_L.png | 1 | 1 | 1 | 1 | 0.606666666 |  |  |
| AD18\_R.png | 1 | 0 | 1 | 1 | 0.182666666 |  |  |
| AD19\_L.png | 1 | 1 | 1 | 1 | 0.029333334 |  |  |
| AD19\_R.png | 1 | 1 | 1 | 1 | 0.397142858 |  |  |
| AD1\_L.png | 1 | 1 | 1 | 1 | 0.624285714 |  |  |
| AD20\_L.png | 1 | 0 | 1 | 1 | 0.4225 |  |  |
| AD21\_L.png | 1 | 1 | 0 | 1 | 0.021333334 |  |  |
| AD21\_R.png | 1 | 1 | 1 | 1 | 0.22 |  |  |
| AD22\_L.png | 1 | 1 | 0 | 1 | 0.166666666 |  |  |
| AD23\_L.png | 1 | 1 | 1 | 1 | 0.716 |  |  |
| AD24\_L.png | 1 | 1 | 1 | 1 | 0.28923077 |  |  |
| AD25\_L.png | 1 | 1 | 0 | 1 | 0.004615384 |  |  |
| AD26\_L.png | 1 | 0 | 1 | 0 | 0.708 |  |  |
| AD27\_R.png | 1 | 0 | 1 | 1 | 0.08 |  |  |
| AD28\_R.png | 1 | 1 | 1 | 1 | 0.205333334 |  |  |
| AD29\_R.png | 1 | 1 | 1 | 1 | 0.342857142 |  |  |
| AD2\_L.png | 1 | 1 | 1 | 1 | 0.274285714 |  |  |
| AD30\_R.png | 1 | 1 | 1 | 1 | 0.654285714 |  |  |
| AD31\_R.png | 1 | 1 | 1 | 1 | 0.512 |  |  |
| AD32\_L.png | 1 | 1 | 0 | 0 | 0.227692308 |  |  |
| AD33\_L.png | 1 | 1 | 1 | 1 | 0.574285714 |  |  |
| AD34\_L.png | 1 | 1 | 1 | 1 | 0.745714286 |  |  |
| AD34\_R.png | 1 | 0 | 1 | 1 | 0.14 |  |  |
| AD36\_R.png | 1 | 0 | 1 | 1 | 0.422666666 |  |  |
| AD3\_L.png | 1 | 0 | 1 | 1 | 0.3025 |  |  |
| AD48\_R.png | 1 | 0 | 0 | 1 | 0.06 |  |  |
| AD4\_L.png | 1 | 1 | 1 | 1 | 0.048571428 |  |  |
| AD4\_R.png | 1 | 1 | 1 | 1 | 0.27 |  |  |
| AD55\_R.png | 1 | 0 | 0 | 0 | 0.138666666 |  |  |
| AD5\_L.png | 1 | 0 | 0 | 0 | 0.152307692 |  |  |
| AD5\_R.png | 1 | 1 | 1 | 1 | 0.132857142 |  |  |
| AD6\_R.png | 1 | 1 | 0 | 1 | 0.482857142 |  |  |
| AD7\_R.png | 1 | 0 | 1 | 1 | 0.118666666 |  |  |
| AD8\_R.png | 1 | 1 | 0 | 1 | 0.056 |  |  |
| AD9\_R.png | 1 | 0 | 0 | 0 | 0.217142858 |  |  |

Supplementary Figure 2: Table comparing ground truth, classification before gist input, classification after gist input, and radiologist gist response. Color coded to show instances errors were introduced or resolved by the model or radiologists gist input.