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 0.899 - higher than the radiologists or the model without gist input alone.

Gist processing is the visual perception skill humans use to quickly assess semantic information from a scene through only a glance that lasts a fraction of a second. This processing allows us to intelligently allocate 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]. Other experiments and contests have pushed deep learning models to perform well at classifying mammograms [11-12]. 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. We utilize the knowledge that the radiologists already have to speed up and improve results.

We hypothesize that by using radiologist gist response within a classifier used to screen full-field mammograms we will see improved results over just the radiologist or the machine without the radiologist input. Our approach utilizes radiologist gist readings with features from a deep convolutional neural network (CNN) used to train a classifier to screen mammograms for subtle abnormalities. Four pre-processing approaches are tested as inputs to the deep CNNs. We test both InceptionV4 and VGG19 features, with LightGBM(LBGM) or a linear SVM as a classifier, giving us 16 classifiers total. We test each of these classifiers with and without radiologist gist responses appended to the feature vector for input. 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]. 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 models when adding the radiologist gist response into the classifier’s input, with these new classifiers even outperforming radiologists in some cases.

\* jwolfe@bwh.harvard.edu

1: Ohio State University

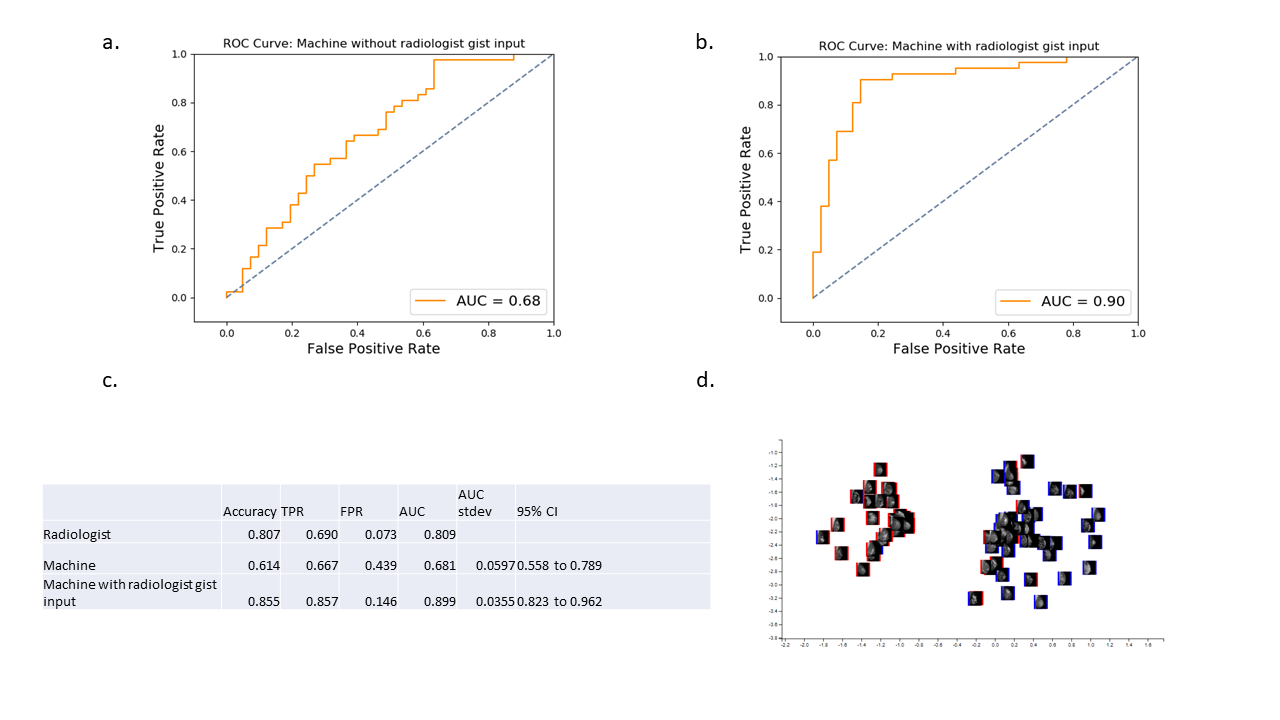
2: University of York

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. When testing the 16 classifiers without using radiologist gist response as part of the input feature vector to the classifier, all resulting AUCs are worse than the radiologists except for one. When all original left breast images are mirrored in pre-processing and used as inputs into VGG19 and classified with a linear SVM, we have an AUC of 0.828 (95% CI 0.739 to 0.908). On average, classifiers not using radiologist gist response in the input vector have an AUC of 0.656.

After appending the radiologist gist response to the feature vectors, the average AUC of the 16 classifiers increases to 0.809. On Average, each classifier improves its AUC by 0.153, reduces false positive rate (FPR) by 0.183, improves true positive rate (TPR) by 0.095, and improves accuracy by 0.139. In all cases, AUC, accuracy, TPR, and FPR improve when adding radiologist gist response to the input vector. In 10 of the 16 classifiers that use radiologist gist response in the input vector, we observe an AUC higher than the radiologists. One result reports an AUC of 0.899. This was achieved with pre-processing the left breast images to be mirrored before input to InceptionV4, and then using LGBM for classification. We further analyse this classifier’s results before and after adding the radiologist gist response to the input vector.



1. All figures show data from the LGBM classifier that used left breast mirrored pre-processing features from InceptionV4. **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 radiologist gist response and the classifier with and without radiologist gist response appended to the input vector. **d.** t-SNE visualization of the input images, using the classifier’s output as well as a 50-dimensional PCA of the deep feature vector for each image from InceptionV4 appended. The color on the left is what the classifier classified the image as, and the color on the right is the ground truth. Blue is normal, red is abnormal.

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 made by the classifier without gist response input 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.

In this case, we see an improvement over the classifier not utilizing radiologist gist response, improving AUC from 0.681(95% CI 0.558 to 0.789) to 0.899(95% CI 0.823 to 0.962). A Levene Test for equality of variances between the classifier’s AUC before gist input, classifier AUC after gist input, and radiologist AUC fails (F(2, 2997) = 4.58, p=0.000), so we use Welch’s F test. There was a statistically significant difference between groups as determined by the one-way Welch’s F test (F(2, 1930) = 5117, p=0.000). A Games-Howell post hoc test revealed that the AUC of the classifier without gist input was statistically significantly lower than the radiologists gist response(0.679 +- 0.0597, p=0.000), and that the classifier with gist input was statistically significantly higher than the radiologists gist response(0.897 +- 0.0396, p=0.000).

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. This is further supported by the fact that 10 of the 16 models using gist input appended to the feature vectors had higher AUC than radiologist gist response, indicating that the inputs must capture different signals within the mammograms.

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 deep features 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 CNNs and machine learning. The combination of the parts is better than either solution alone, as each input captures different features. 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 and may pick up on different signals than a deep neural network might.

# 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. Bootstrapping on the radiologist classifications is done with 1000 samples to estimate the variance of the radiologist gist AUC. A “confidence score” for each image is calculated based on the average response’s difference from 50, then normalized.

It has been shown that cropping out pectoral muscle during pre-processing may be beneficial for classifiers when screening full-field mammograms for breast cancer. Due to our small dataset, we test four pre-processing techniques which involve mirroring and cropping. These are as follows: (1) No crop – do not crop the pectoral muscle, and use the original full-field mammograms; (2) No crop, same direction – do not crop the pectoral muscle, but mirror all left breast mammograms, such that all mammograms are on the same side. We expect this reduces the abstraction that the classifier needs to learn during training, which it might not achieve with a dataset of 83; (3) Crop – crop pectoral muscle out of the original mammograms by setting those pixel values to black; (4) Crop, same direction – crop the pectoral muscle out of the original mammograms, and then mirror all left breast images for the same reasons as (2). We implement the cropping algorithm introduced in [8], and visually inspect results to correct over and under cropping.

To generate deep features for the input images, we test two deep CNNs, VGG19 and InceptionV4. Both are pre-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 both models. We take the values from nodes located in the first fully connected layer after the convolutional layers are used capture features about a photo fed through the network from InceptionV4, and use the values from the ReLu6 layer in VGG19. Thus, we have 1536 features from InceptionV4 and 4096 features from VGG19. This creates a feature vector, which is used as input into a supervised learning algorithm. We test Microsoft’s LightGBM and a linear support vector machine (SVM) 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. For the linear SVM, C=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. After the pre-processing step above, images were downscaled to 299x299 pixels to fit the pre-trained InceptionV4 model properly via a function in pretrainedmodels, or 224x224 to fit the pre-trained VGG19 model. Image pixel values are normalized as they were when InceptionV4 and VGG19 were trained. Leave-one-out(LOO) training and testing is used to generate 83 out-of-fold(OOF) predictions for each image in the set. Predictions will be 0 if classified as normal, and 1 if classified as malignant. With our 4 pre-processing methods, 2 deep CNNs, and 2 classifiers, we have 16 models for classification that don’t use radiologist gist response input.

The above process is repeated to create another 16 models that incorporate the radiologist gist input by appending the radiologist gist classification (0 or 1) and the calculated radiologist gist confidence (0.0-1.0), each multiplied by 100. This creates a new feature vector of length 1538 for each image when testing InceptionV4 feature vectors, and a feature vector length of 4098 when testing VGG19 feature vectors. Bootstrapping 1000 AUC samples on the final predictions for each model is used to estimate the variance of the area under the receiver operating characteristic (AUC) for that model. A 95% confidence interval is constructed using the middle 95% of samples from the bootstrapping set to account for the any skew in the samples. Cohen’s d is calculated using the mean difference divided by the standard deviation of the model.Testing significance between radiologist confidence subsets is done using two-sample t-tests.

References

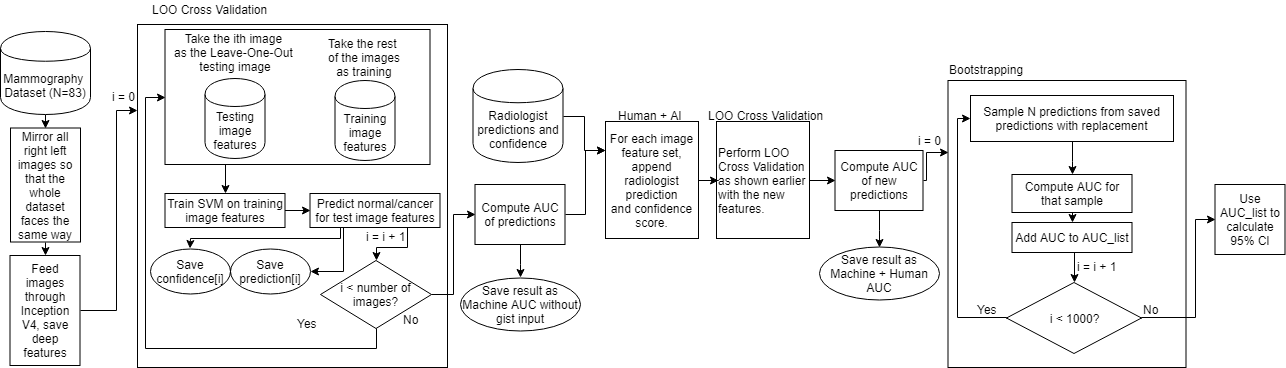
1. M. Kalager, M. Zelen, F. Langmark, and H. O. Adami. New England Journal of Medicine, 363(13):1203–1210, 2010.
2. B. Q. Huynh, H. Li, and M. L. Giger. Journal of Medical Imaging, 3(3):034501–034501, 2016.
3. Giger M. L., Boone J., Chan H.,. Med. Phys. 35(12), 5799–5820 (2008).
4. Hadjiiski L., Sahiner 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.
11. Ribli, Dezső, Horváth, Anna, Unger, Zsuzsa, Pollner, Péter, Csabai, István. Scientific Reports Volume 8 Article Number 4165 (2018)
12. Tsochatzidis, L.; Costaridou, L.; Pratikakis, I. J. Imaging 2019, 5, 37.
13. Abdelhafiz, D., Yang, C., Ammar, R. et al. BMC Bioinformatics (2019) 20(Suppl 11): 281. <https://doi.org/10.1186/s12859-019-2823-4>
14. Ren, Shaoqing, et al. Advances in neural information processing systems. 2015.
15. Qin, Zhuwei, et al. arXiv preprint arXiv:1804.11191 (2018).
16. Szegedy, Christian, et al. Thirty-First AAAI Conference on Artificial Intelligence. 2017.
17. Adam Yala, Constance Lehman, Tal Schuster, Tally Portnoi, Regina Barzilay. Radiology, Volume 292, No. 1 (2019)
18. Arkadiusz Sitek, Jeremy M. Wolfe. Radiology. https://doi.org/10.1148/radiol.2019190791

# 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 radiologist’s gist input.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Accuracy | TPR | FPR | AUC | AUC stdev | 95% CI |
| Radiologist | 0.807228916 | 0.69047619 | 0.073170732 | 0.808652729 |  |  |
| LGBM |  |  |  |  |  |  |
| Machine(1) | 0.530120482 | 0.547619048 | 0.487804878 | 0.554006969 | 0.063264811 | 0.427 to 0.674 |
| Machine+Radiologist(1) | 0.626506024 | 0.666666667 | 0.414634146 | 0.739837398 | 0.052181566 | 0.632 to 0.838 |
| Machine(2) | 0.638554217 | 0.761904762 | 0.487804878 | 0.658536585 | 0.060684685 | 0.525 to 0.773 |
| Machine+Radiologist(2) | 0.746987952 | 0.857142857 | 0.365853659 | 0.801393728 | 0.048627639 | 0.696 to 0.8854 |
| Machine(3) | 0.530120482 | 0.595238095 | 0.536585366 | 0.515679443 | 0.062875448 | 0.388 to 0.640 |
| Machine+Radiologist(3) | 0.638554217 | 0.666666667 | 0.390243902 | 0.724157956 | 0.05425109 | 0.615 to 0.827 |
| Machine(4) | 0.626506024 | 0.619047619 | 0.365853659 | 0.646341463 | 0.059974977 | 0.526 to 0.761 |
| Machine+Radiologist(4) | 0.662650602 | 0.642857143 | 0.317073171 | 0.763356562 | 0.051471411 | 0.655 to 0.853 |
| SVM |  |  |  |  |  |  |
| Machine(1) | 0.65060241 | 0.666666667 | 0.365853659 | 0.700929152 | 0.05838642 | 0.580 to 0.814 |
| Machine+Radiologist(1) | 0.78313253 | 0.761904762 | 0.195121951 | 0.850174216 | 0.042445374 | 0.756 to 0.925 |
| Machine(2) | 0.722891566 | 0.714285714 | 0.268292683 | 0.828106852 | 0.044197804 | 0.739 to 0.908 |
| Machine+Radiologist(2) | 0.819277108 | 0.738095238 | 0.097560976 | 0.861207898 | 0.040536965 | 0.776 to 0.931 |
| Machine(3) | 0.530120482 | 0.571428571 | 0.512195122 | 0.553426249 | 0.064682387 | 0.427 to 0.681 |
| Machine+Radiologist(3) | 0.759036145 | 0.761904762 | 0.243902439 | 0.806620209 | 0.047687339 | 0.702 to 0.891 |
| Machine(4) | 0.65060241 | 0.642857143 | 0.341463415 | 0.680023229 | 0.056282573 | 0.570 to 0.789 |
| Machine+Radiologist(4) | 0.78313253 | 0.69047619 | 0.12195122 | 0.81358885 | 0.047490594 | 0.717 to 0.898 |

Supplementary Figure 3: Table of results using VGG19 deep features. The number in parenthesis is the pre-processing technique used for input. “Machine+Radiologist” means that the radiologist gist response was appended to the input feature vectors. The top 8 results used the LGBM classifier, and the bottom 8 used an SVM.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Accuracy | TPR | FPR | AUC | AUC stdev | 95% CI |
| Radiologist | 0.807228916 | 0.69047619 | 0.073170732 | 0.808652729 |  |  |
| LGBM |  |  |  |  |  |  |
| Machine(1) | 0.686746988 | 0.714285714 | 0.341463415 | 0.743321719 | 0.05539068 | 0.628 to 0.843 |
| Machine+Radiologist(1) | 0.771084337 | 0.785714286 | 0.243902439 | 0.825783972 | 0.046795534 | 0.726 to 0.910 |
| Machine(2) | 0.614457831 | 0.666666667 | 0.43902439 | 0.681184669 | 0.059718796 | 0.558 to 0.789 |
| Machine+Radiologist(2) | 0.855421687 | 0.857142857 | 0.146341463 | 0.898954704 | 0.035593996 | 0.823 to 0.962 |
| Machine(3) | 0.56626506 | 0.69047619 | 0.56097561 | 0.62485482 | 0.061557764 | 0.501 to 0.739 |
| Machine+Radiologist(3) | 0.662650602 | 0.714285714 | 0.390243902 | 0.701509872 | 0.056233201 | 0.591to 0.804 |
| Machine(4) | 0.626506024 | 0.69047619 | 0.43902439 | 0.675958188 | 0.058833446 | 0.555 to 0.784 |
| Machine+Radiologist(4) | 0.734939759 | 0.714285714 | 0.243902439 | 0.81416957 | 0.045457731 | 0.723 to 0.899 |
| SVM |  |  |  |  |  |  |
| Machine(1) | 0.578313253 | 0.642857143 | 0.487804878 | 0.669570267 | 0.058234816 | 0.554 to 0.781 |
| Machine+Radiologist(1) | 0.795180723 | 0.785714286 | 0.195121951 | 0.831591173 | 0.047535313 | 0.733 to 0.918 |
| Machine(2) | 0.65060241 | 0.642857143 | 0.341463415 | 0.638792102 | 0.060675892 | 0.513 to 0.755 |
| Machine+Radiologist(2) | 0.795180723 | 0.785714286 | 0.195121951 | 0.835075494 | 0.045916273 | 0.742 to 0.921 |
| Machine(3) | 0.602409639 | 0.642857143 | 0.43902439 | 0.665505226 | 0.059838727 | 0.542 to 0.780 |
| Machine+Radiologist(3) | 0.819277108 | 0.80952381 | 0.170731707 | 0.839140534 | 0.046419442 | 0.740 to 0.922 |
| Machine(4) | 0.614457831 | 0.666666667 | 0.43902439 | 0.665505226 | 0.060694873 | 0.535 to 0.774 |
| Machine+Radiologist(4) | 0.78313253 | 0.761904762 | 0.195121951 | 0.844367015 | 0.044918804 | 0.755 to 0.929 |

Supplementary Figure 4: Table of results using InceptionV4 deep features. The number in parenthesis is the pre-processing technique used for input. “Machine+Radiologist” means that the radiologist gist response was appended to the input feature vectors. The top 8 results used the LGBM classifier, and the bottom 8 used an SVM.