Interpretable Deep Learning Models for Better Clinician-AI Communication in Clinical Mammography

Alina Jade Barnett[a], Vaibhav Sharma[a], Neel Gajjar[a], Jerry Fang[a], Fides Regina Schwartz[b], Chaofan Chen[c], Joseph Y. Lo[b,d,e], and Cynthia Rudin[a,e]

[a]Department of Computer Science, Duke University, Durham, USA
[b]Department of Radiology, Duke University, Durham, USA
[c]Department of Computer Science, University of Maine, Orono, USA
[d]Department of Biomedical Engineering, Duke University, Durham, USA
[e]Department of Electrical and Computer Engineering, Duke University, Durham, USA

ABSTRACT

There is increasing interest in using deep learning and computer vision to help guide clinical decisions, such as whether to order a biopsy based on a mammogram. Existing networks are typically black box, unable to explain how they make their predictions. We present an interpretable deep-learning network which explains its predictions in terms of BI-RADS features mass shape and mass margin. Our model predicts mass margin and mass shape, then uses the logits from those interpretable models to predict malignancy, also using an interpretable model. The interpretable mass margin model explains its predictions using a prototypical parts model. The interpretable mass shape model predicts segmentations, fits an ellipse, then determines shape based on the goodness of fit and eccentricity of the fitted ellipse. While including mass shape logits in the malignancy prediction model did not improve performance, we present this technique as part of a framework for better clinician-AI communication.

Keywords: Interpretability, Deep Learning, Mammography, Masses, Cancer

1. INTRODUCTION & PURPOSE

Deep learning is now pervasive in radiology, but even FDA-approved models are black boxes that we cannot fully understand or trust. However, accurate deep learning models for radiology do not need to be black boxes. Saliency methods have been introduced to remedy this, but while saliency methods such as GradCAM are commonly used, there are known issues[1–3] in that their results are not necessarily interpretable, or even correct. In contrast, several newer works show it is possible for neural networks to explain their reasoning processes in an interpretable way humans can understand, with explanations more detailed than those given by saliency alone.[4,5] Such interpretable models are challenging to build, but provide insights that cannot be achieved otherwise. Previously, Barnett et al.[6] constructed a model called "IAIA-BL," Interpretable AI Algorithm for Breast Lesions. This model predicts the malignancy of a breast lesion by using the logits of an interpretable mass margin classification model. However, in a report written by a radiologist, a breast mass lesion must be described not only by its margin but also by its shape to conform with BI-RADS requirements. A mass may be classified as one of three shapes: irregular, round, or oval; as shown in Figure 2. Classification of mass shape with uninterpretable CNNs has been previously studied.[7,8] In this paper, we present an interpretable mass shape model. By combining our interpretable mass shape model with IAIA-BL's mass margin model, our technique not only accurately predicts biopsy outcomes but also empowers radiologists to understand the model's recommendations.[9,10]

2. NEW WORK

An interpretable mass shape model is introduced in this submission. Previous submissions of IAIA-BL have demonstrated its effectiveness as an interpretable model for malignancy prediction, however, this has been limited to predictions using mass margin.[6] This submission explores the use of interpretable mass shape alongside interpretable mass margin to predict malignancy.

3. METHODS

3.1 Data

Under Duke University Health IRB Pro00012010, we collected 1136 images from 484 patients who received mammograms and biopsies at Duke University Health Systems between 2008 and 2018. Data was labelled as malignant or benign based on the histopathology report. The images were further labelled with mass margin and mass shape by one fellowship-trained breast imaging radiologist. We divided the data into training images (73%), validation images (12%), and testing images (15%) such that there was no patient overlap between sets. Figure 3 shows the distribution of mass margin, mass shape and malignancy across the data split.

3.2 Segmentation maps

The segmentation network was trained on the publicly available CBIS-DDSM dataset.[11] We used a UNet with a base architecture of VGG-16, pretrained on ImageNet. To obtain the predicted binary segmentation maps, we selected the largest contiguous region above the threshold value of 0.2 (where the UNet outputs values between 0 and 1).

3.3 From segmentation map to shape class

Refer to Figure 1. Given an enclosed shape R whose edge traces the edge of the mass, we find the ellipse C that best fits R. A is the set of all pixels that are enclosed by exactly one of R or C, as shown in Figure 1d.

We classify the shape of a mass as follows:

1. If $Area(A) \geq \tau Area(C)$, then the mass is classified as "irregular." When the area of A is low relative to the area of C, the mass boundary is a close fit to the ellipse. For masses with a high relative area of A, the mass edge cannot be fit closely to an ellipse.

2. If $Area(A) < \tau Area(C)$ and the eccentricity of ellipse $e_C \leq \nu$ (circles or short ovals), the mass is classified as "round." The eccentricity of an ellipse is the ratio between the length of the semi-major axis and the length of the semi-minor axis; refer to Figure 4 for examples of ellipses with varying eccentricity.

3. If $Area(A) < \tau Area(C)$ and the eccentricity of the ellipse $e_C > \nu$ (elongated oval), the mass is classified as "oval."

Parameters τ and ν are learned from data labelled by radiologists. The best τ and ν values were selected by taking the result with highest average recall between the model and the annotator on the institutional training data, rejecting any model whose predictions do not span all classes.

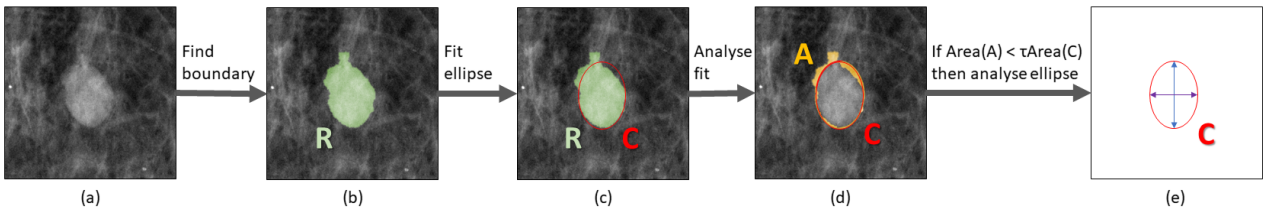

Figure 1: (a) The lesion to be analysed. (b) The lesion is segmented by green shape R. (c) Ellipse C is fitted to the shape given by R. (d) A shows the difference between the areas enclosed by R and C. (e) If the mass is not irregular (i.e., $Area(A) < \tau Area(C)$), then we consider the eccentricity of fitted ellipse C to determine whether the mass is round or oval.

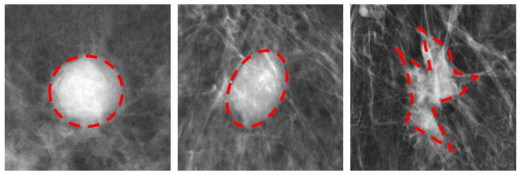

Figure 2: The three mass shapes from the BI-RADS lexicon.

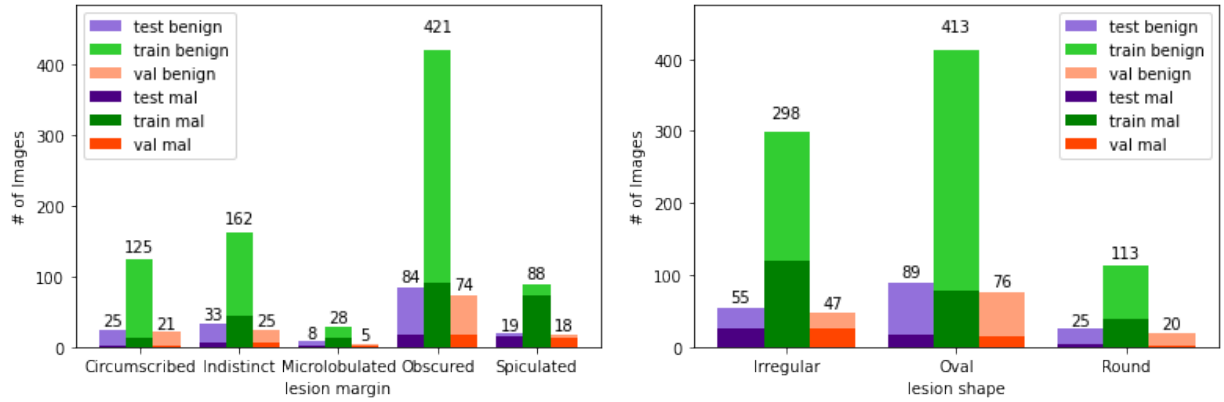

Figure 3: Shows the distribution of (A) mass margin, (B) mass shape and malignancy for both across the data split.

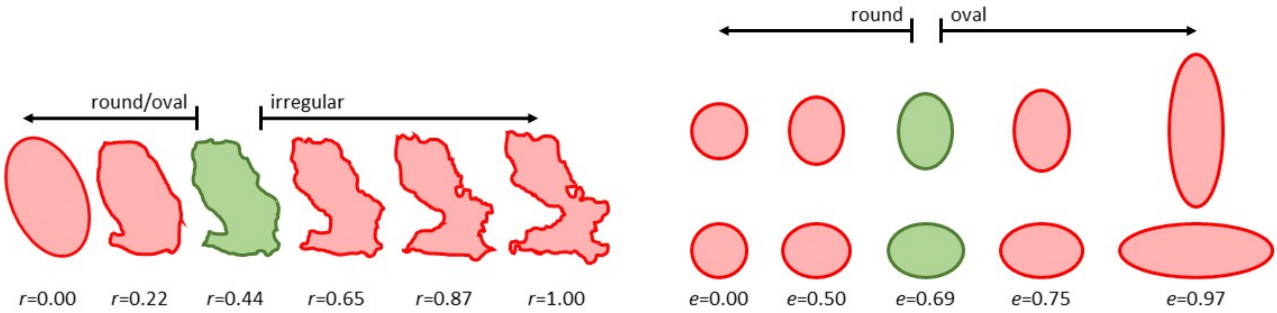

Figure 4: Left: As the ratio $r = Area(A)/Area(C)$ increases, the mass is more likely to be classified as irregular. The decision boundary between irregular and other classes is $r = 0.44$. Right: As eccentricity increases, ellipses become less circular and more elongated. The decision boundary between round and oval is $e = 0.69$.

Model input	Train AUROC	Test AUROC
Margin labels	0.83 [0.78, 0.88]	0.86 [0.74, 0.95]
Shape labels	0.69 [0.64, 0.74]	0.71 [0.57, 0.84]
Margin labels & shape labels	0.83 [0.78, 0.88]	0.83 [0.70, 0.94]
Margin predictions (IAIA-BL)	0.83 [0.78, 0.88]	0.86 [0.76, 0.97]
Margin predictions & shape labels	0.83 [0.78, 0.88]	0.85 [0.74, 0.96]
Margin predictions & shape predictions (ours)	0.84 [0.79, 0.89]	0.85 [0.73, 0.96]

Table 1: The final stage of IAIA-BL uses a linear model to combine evidence from mass margin prediction scores. We present results for malignancy prediction given different model inputs. We use cross-validated logistic regression to train the linear model on the training data. 95% confidence intervals are determined using the Delong method.[12]

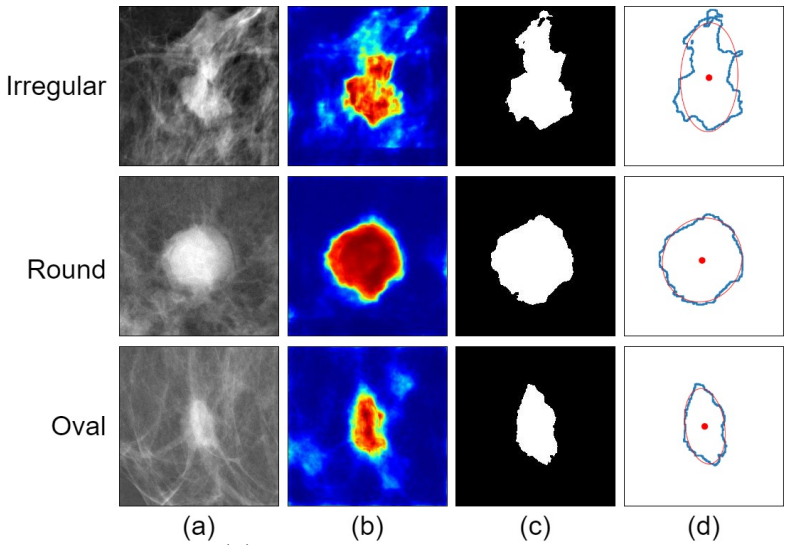

Figure 5: (a) The mass to be classified. (b) Output from the uninterpretable UNet, where red indicates a high predicted probability of being part of the mass and blue indicates a low predicted probability of being part of the mass. (c) Taking the threshold of the UNet predictions from b, then selecting the largest contiguous region. (d) The border and ellipse fit.

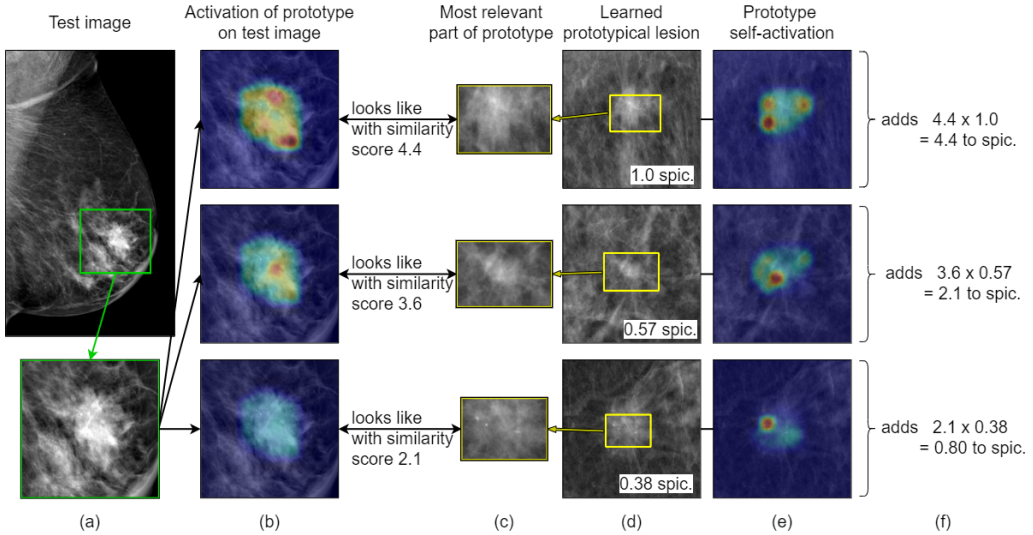

Figure 6: The interpretable network explains why the mass margin is predicted to be spiculated. Reproduced with permission from Barnett et al. (2021)[6]

4. RESULTS

The trained UNet achieves an intersection over union (IOU) of 0.76 on the test set of CBIS-DDSM and 0.73 on the training set of CBIS-DDSM. We cannot report the IOU for the UNet on our institutional data as we do not have ground truth segmentation maps. For our institutional data, we also selected a subset of 273 cases that were judged qualitatively to have the 'best' segmentation maps. With $\tau = 0.44$ and $\nu = 0.69$, the shape prediction model achieves an average recall of 36% on the institutional training set and 33% on the institutional test set. On the subset of segmentation maps judged 'best,' the shape prediction model achieves an average recall of 40%. Several sample explanations from our method are presented in Figure 5. The interpretable mass margin prediction model achieves AUROC of 0.92 on the task of classifying mass margin, with explanations as in Figure 6. We will refer to output of this model as "mass margin prediction scores."

In the previous work, the final stage of IAIA-BL uses a linear model to combine evidence from only the mass margin prediction scores to predict malignancy. When we add shape prediction information as an additional input to the malignancy prediction model, there is no statistically significantly improvement in accuracy, as reported in Table 1. When including ground truth shape labels in the malignancy model, we achieve the same result. This result could have arisen because mass margin is a strong indicator of malignancy, and including mass shape did not contribute additional predictive power beyond that of mass margin. We know that mass shape is correlated with mass margin, so it is possible that adding mass shape improves malignancy prediction, but our dataset is too small to capture the small improvement gained when adding a closely correlated variable. We also investigated the possibility that the mass margin prediction scores contain information about the mass shape, but found that this does not appear to be the case. We trained three linear models to predict mass shape; one using mass margin prediction scores, one using mass margin labels, and one using both mass margin labels and mass margin prediction scores. We found no improvement in mass shape prediction when using mass margin prediction scores compared to using mass margin labels alone.

5. CONCLUSIONS

In any mammographic report that describes a mass, mass shape is a *mandatory* descriptor as part of the BI-RADS classification. We recommend including mass shape as a way to improve clinician-AI communication. We increased the interpretability of the IAIA-BL model by including the BI-RADS feature mass shape. However, for this dataset, including the mass shape did not improve malignancy prediction. There are several possible reasons for this unintuitive result, including the possibility of our dataset not generalizing more broadly or our shape classification not being sufficiently accurate. Follow-up work will be needed to determine which of these is true. However, given that shape must currently be reported in radiologists' reports, our work could easily be useful in assisting with this task.

ACKNOWLEDGMENTS

This study was supported in part by MIT Lincoln Laboratory, Duke TRIPODS, Duke MEDx: High-Risk High-Impact Challenge, and the Duke Incubation Fund. We would like to acknowledge breast radiologists Michael Taylor-Cho MD, Lars Grimm MD, Connie Kim MD, and Sora Yoon MD, who annotated the dataset used in this paper. This study was supported in part by NIH/NCI U01-CA214183 and U2C-CA233254.

REFERENCES

[1] Nguyen, G., Kim, D., and Nguyen, A., "The effectiveness of feature attribution methods and its correlation with automatic evaluation scores," *Neural Information Processing Systems (NeurIPS)* (2021).

[2] Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., and Kim, B., "Sanity checks for saliency maps," *Neural Information Processing Systems (NeurIPS)* (2018).

[3] Draelos, R. L. and Carin, L., "HiResCAM: Faithful location representation in visual attention for explainable 3d medical image classification," *arXiv preprint arXiv:2011.08891* (2020).

[4] Chen, C., Li, O., Tao, C., Barnett, A., Su, J., and Rudin, C., "This looks like that: Deep learning for interpretable image recognition," *Neural Information Processing Systems (NeurIPS)* (2019).

[5] Barnett, A. J., Schwartz, F. R., Tao, C., Chen, C., Ren, Y., Lo, J. Y., and Rudin, C., "A case-based interpretable deep learning model for classification of mass lesions in digital mammography," *Nature Machine Intelligence* (2021).

[6] Barnett, A. J., Schwartz, F. R., Tao, C., Chen, C., Ren, Y., Lo, J. Y., and Rudin, C., "Interpretable mammographic image classification using cased-based reasoning and deep learning," *Deep Learning, Case-Based Reasoning, and AutoML: Present and Future Synergies, Workshop of International Joint Conferences on Artificial Intelligence Organization* (2021).

[7] Singh, V. K., Romani, S., Rashwan, H. A., Akram, F., Pandey, N., Sarker, M. M. K., Abdulwahab, S., Torrents-Barrena, J., Saleh, A., Arquez, M., et al., "Conditional generative adversarial and convolutional networks for X-ray breast mass segmentation and shape classification," *International Conference on Medical Image Computing and Computer-Assisted Intervention* , 833–840 (2018).

[8] Kim, S. T., Lee, H., Kim, H. G., and Ro, Y. M., "ICADx: interpretable computer aided diagnosis of breast masses," *Medical Imaging 2018: Computer-Aided Diagnosis* **10575**, 1057522 (2018).

[9] Edwards, B., "FDA Guidance on Clinical Decision Support: Peering Inside the Black Box of Algorithmic Intelligence." https://www.chilmarkresearch.com/fda-guidance-clinical-decision-support/ (December 2017). Online; accessed March 13, 2018.

[10] Soffer, S., Ben-Cohen, A., Shimon, O., Amitai, M. M., Greenspan, H., and Klang, E., "Convolutional neural networks for radiologic images: a radiologist's guide," *Radiology* **290**(3), 590–606 (2019).

[11] Lee, R. S., Gimenez, F., Hoogi, A., and Rubin, D., "Curated breast imaging subset of DDSM," *The cancer imaging archive* **8**, 2016 (2016).

[12] DeLong, E. R., DeLong, D. M., and Clarke-Pearson, D. L., "Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach," *Biometrics* , 837–845 (1988).