nature machine intelligence

ARTICLES

https://doi.org/10.1038/s42256-021-00423-x

Check for updates

A case-based interpretable deep learning model for classification of mass lesions in digital mammography

Alina Jade Barnett[1]✉, Fides Regina Schwartz[2], Chaofan Tao[1], Chaofan Chen[3], Yinhao Ren[4], Joseph Y. Lo[2,4,5] and Cynthia Rudin[1,5,6,7]

Interpretability in machine learning models is important in high-stakes decisions such as whether to order a biopsy based on a mammographic exam. Mammography poses important challenges that are not present in other computer vision tasks: datasets are small, confounding information is present and it can be difficult even for a radiologist to decide between watchful waiting and biopsy based on a mammogram alone. In this work we present a framework for interpretable machine learning-based mammography. In addition to predicting whether a lesion is malignant or benign, our work aims to follow the reasoning processes of radiologists in detecting clinically relevant semantic features of each image, such as the characteristics of the mass margins. The framework includes a novel interpretable neural network algorithm that uses case-based reasoning for mammography. Our algorithm can incorporate a combination of data with whole image labelling and data with pixel-wise annotations, leading to better accuracy and interpretability even with a small number of images. Our interpretable models are able to highlight the classification-relevant parts of the image, whereas other methods highlight healthy tissue and confounding information. Our models are decision aids—rather than decision makers—and aim for better overall human–machine collaboration. We do not observe a loss in mass margin classification accuracy over a black box neural network trained on the same data.

Artificial intelligence is revolutionizing radiology. Consider mammography, which aims to detect breast cancer, a leading cause of death in the USA[1]. Conventional machine learning is currently used for computer-aided detection (that is, lesion versus no lesion), but future approaches need to be able to assist with harder tasks to make a greater clinical contribution, for example: 'should the patient get a biopsy for that lesion?'

Despite the hope of computer-aided radiology for mammography, there are serious concerns with current methods, including confounding. Confounding occurs when the predictive model is using incorrect information or reasoning to make a decision, even if the decision is correct. In previous studies, researchers created models that seemed to perform well on their test sets, yet on further inspection, based their decisions on confounding information (for example, type of equipment) rather than medical information[2–4]. This problem is exacerbated by the fact that there are few publicly available mammography datasets, so many models are trained on relatively few cases and the community lacks datasets to externally validate these models. Furthermore, not all studies incorporate domain expertise to inform which factors should be included in the model building phase. To ensure clinical acceptance, an AI tool will need to provide its reasoning process to its human radiologist collaborators to be a useful aide in these difficult and high-stakes decision-making processes[5,6].

The reasoning process of any model would ideally be similar to that of an actual radiologist, who looks at specific aspects of the image that are known to be important, based on the physiology of how lesions develop within breast tissue. If this reasoning process were correct, it would lead to a higher chance that: (1) the model could generalize beyond its finite training set; (2) the model's reasoning process could be useful information for doctors, even if its prediction is sometimes incorrect; (3) it would be much easier to troubleshoot or evaluate trustworthiness of the model, as it is not a black box; (4) the model's reasoning and reporting process could align with the structured lexicon that radiologists use to report results, such as the breast imaging-reporting and data system (BI-RADS)[7] for breast cancer and other similar lexicons from the American College of Radiology.

Thus, unlike existing black-box systems that aim to replace a doctor[8], we aim to create an interpretable AI algorithm for breast lesions (abbreviated IAIA-BL) whose explicit reasoning can be understood and verified by a medical practitioner. Our novel deep learning architecture enables IAIA-BL to provide an explanation that shows the underlying decision-making process for each case. Figure 1c shows an example of how this works: the algorithm highlights parts of the image, explains that it considers these parts of images similar to prototypical cases it has seen before, and provides a score for the probability of the specific diagnosis (breast mass with mostly circumscribed margin) for this image as well as the likelihood of malignancy. IAIA-BL provides the radiologist with the means not to simply trust the AI but to check its output for plausibility, and overrule it when necessary. Our approach, with its inherently interpretable reasoning process, contrasts directly with previous work that relied on post-hoc explanation techniques to

[1]Department of Computer Science, Duke University, Durham, NC, USA. [2]Department of Radiology, Duke University, Durham, NC, USA. [3]School of Computing and Information Science, University of Maine, Orono, ME, USA. [4]Department of Biomedical Engineering, Duke University, Durham, NC, USA. [5]Department of Electrical and Computer Engineering, Duke University, Durham, NC, USA. [6]Department of Statistical Science, Duke University, Durham, NC, USA. [7]Department of Biostatistics & Bioinformatics, Duke University, Durham, NC, USA. ✉e-mail: alina.barnett@duke.edu

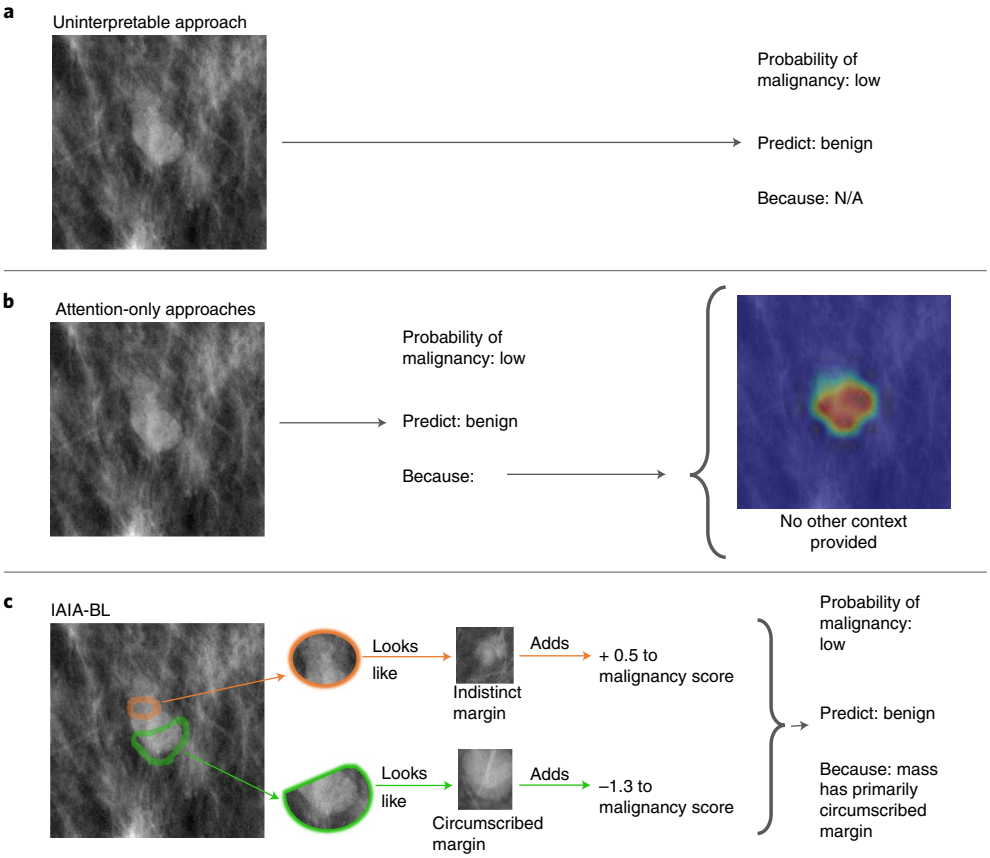


Fig. 1 | An overview of IAIA-BL compared to other approaches. a, Uninterpretable approaches give no explanations for their output. **b**, Other interpretable or explainable approaches might point out which regions are used for decision making, but provide no information about what attributes of the region are important for classification decisions. **c**, IAIA-BL provides an explanation framework that localizes relevant areas, associates the relevant area with a specific medical feature and uses only the explained evidence to make a prediction.

explain a trained black-box model, or work that relies on attention mechanisms to highlight the parts of an input image on which the model prediction is based. The reasoning process the network explains to the humans is the reasoning process it is using to understand the image itself.

The main contributions of our paper are as follows:

- We developed the first inherently interpretable machine-learning-based system for medical imaging that goes beyond simple attention in its explanations. Our system, IAIA-BL, makes predictions for mammographic breast masses by comparing test mammograms with prototypical images of various mass margin types.
- We developed a novel training scheme for our IAIA-BL, which allows it to incorporate past knowledge in the form of fine-grained expert image annotations. Using only a small number of finely annotated training data and imposing a novel fine-annotation loss on those data, IAIA-BL learns medically relevant prototypes, effectively addresses aspects of confounding issues in medical machine learning models, and sets our IAIA-BL apart from the ProtoPNet presented in ref. [9] and other past works.
- We developed a framework for machine learning-based mammography interpretation in line with the goals of radiologists: in addition to predicting whether a lesion is malignant or benign, our work aims to follow the reasoning processes of radiologists in detecting specific aspects of each image, such as the characteristics of the mass margins.

Related work

Background on computer-aided detection/diagnosis in mammography. Computer-aided detection systems flag suspicious lesions that may prompt the radiologist to recall a patient for additional imaging. Despite widespread clinical adoption, however, an influential study from 2015 found that current systems do not improve diagnoses in clinical practice[10]. More recent deep learning studies based on large numbers of cases have been reported to match or even exceed radiologist performance[8,11–14]. Going beyond lesion detection, computer-aided diagnosis systems provide additional diagnostic information such as classifying the lesion as benign versus malignant[15]. In this study we seek to advance the underlying technology beyond that of previous computer-aided diagnosis approaches.

Background on interpretable machine learning. In spite of their promising performance, deep neural networks are difficult to understand by humans. There are two distinct approaches to address this challenge: (1) designing inherently interpretable networks, whose reasoning process is constrained to be understandable to humans; (2) explaining black box neural networks post-hoc by creating approximations, saliency maps or derivatives. Post-hoc explanations can be problematic, for instance, saliency maps highlight regions of the image, but can be unreliable and misleading as they tend to highlight edges and the highlighted pixels do not show what computations are actually performed[16–18]. We avoid post-hoc solutions in this work. There are several types of approaches in interpretable machine learning, including case-based reasoning (which we use

here), forcing the network to use logical conditions within its last layers (for example, ref. [19]) or disentangling the neural network's latent space[20]. Case-based reasoning models in medicine retrieve existing similar cases to determine how to handle a new case[21–24]. IAIA-BL's framework incorporates the architecture of the case-based inherently interpretable neural network ProtoPNet, described in ref. [9].

Confounding and fine annotation. Neural networks models often use context or confounding information instead of the information that a human would use to solve the same problem in both medical[25] and non-medical applications[26]. The ability of these networks to use context or background information is so powerful that networks trained on images of only the background outperform networks trained on images of only the object to be classified[27,28]. For high-stakes applications in medicine, model decisions must use relevant medical information rather than context or background information. To address this we introduce an attention mechanism, which redirects model attention to a selected part of the input image. Ways to direct model attention include data augmentation[27,29,30], techniques that combat hand-selected confounders[25,31], techniques that combat learned confounders[32] and an approach in which a human critic manually approves the attention map during training[33]. Other techniques that show model attention but do not aim to change it are class activation maps[34], multiattention CNNs[35] and recurrent attention CNNs[36]. Our model uses an attention mechanism to incorporate expert annotations by adding a term to the objective function, which penalizes attention outside of the regions marked as relevant by the radiologist annotator. Mechanically, the method is most similar to that in ref. [31], but differs in that our class-specific attention mechanism asks for different attention from prototypes of different classes.

Data

Our dataset consists of 1,136 digital screening mammogram images from 484 patients at Duke University Health System from 2008 to 2018. These conventional mammography images were collected in compliance with HIPAA under Duke Health IRB Pro00012010 and waiver of informed consent. The average patient age was 56.1 ± 12.3 years. The BI-RADS features of mass shape and mass margin were labelled by one fellowship-trained breast imaging radiologist. The ground truth for malignancy of each mass is the result of definitive histopathology diagnosis.

The 1,136 images consisted of the following mass margins: 125 spiculated, 220 indistinct, 41 microlobulated, 579 obscured and 171 circumscribed. We excluded lesions with microlobulated margins because of the small number of lesions represented. We excluded lesions with obscured margins as this margin class is not a good indicator for classifying a lesion as benign or malignant, but instead usually indicates the need for follow-up imaging. We split each remaining margin class into 73% training, 12% validation and 15% testing, ensuring that within each class there was no patient overlap between the testing set and other sets. All performances are based on the testing set alone ($n = 78$).

We represent the dataset of n training images $\mathbf{x}_i$, with mass-margin (y_i^{margin}) and malignancy (y_i^{mal}) labels, as $D = \{(\mathbf{x}_i, y_i^{\text{margin}}, y_i^{\text{mal}})\}_{i=1}^{n}$; a thirty-image subset $D' \subseteq D$ comes with the radiologist's (fine) annotations of where medically relevant information is in that training image. For a training instance $(\mathbf{x}_i, y_i^{\text{margin}}, y_i^{\text{mal}}) \in D'$, we define a fine-annotation mask $\mathbf{m}_i$, such that $\mathbf{m}_i$ takes the value 0 at those pixels that are marked as 'relevant to mass margin identification,' and takes the value 1 at other pixels. Each fine-annotation mask $\mathbf{m}_i$ has the same spatial dimensions (height and width) as $\mathbf{x}_i$.

Model

Although ProtoPNet[9] works well with bird classification, it could not be directly extended to mammograms due to the problems with confounding, which is made worse by the dearth of data and the difficulty of the overall problem. IAIA-BL overcomes these obstacles through its framework, including incorporating fine-grained labels (Fig. 2), modified modular training and the addition of mult-stage reasoning wherein the model first determines the mass margin feature and uses that information to predict malignancy.

The training of IAIA-BL differs from that of ProtoPNet in three ways: (1) IAIA-BL was trained with a fine-annotation loss which penalizes prototype activations on medically irrelevant regions for the subset of data with fine annotations; (2) IAIA-BL considers the top 5% of the most activated convolutional patches that are closest to each prototype, instead of only the top most activated patch; (3) we include an additional fully connected layer to transform mass margin score ($\hat{\mathbf{y}}^{\text{margin}}$) to the malignancy score ($\hat{y}^{\text{mal}}$), whose training is isolated from the rest of the network.

Results

Performance metrics. We use the area under receiver operator characteristic curve (AUROC) for each of the three mass margin classes as the performance metric for both mass-margin prediction and malignancy prediction. An image-weighted average of these AUROCs measures overall performance; 95% confidence intervals were derived using Delong's method[37,38].

Cohen's κ shows the agreement between our model's predictions and the physician-annotator's labels for the mass margin prediction task. We use Cohen's κ to compare our model's agreement with the agreements between physicians from past studies[39–43]; 95% confidence intervals were derived using non-parametric bootstrap resampling with 5,000 samples each equal to the size of the test set.

Interpretability metric. We designed the interpretability metric activation precision to quantify what proportion of the information used to classify the mass margin comes from the relevant region as marked by the radiologist annotator. Using the notations defined in model training, the activation precision for a single prototype $\mathbf{p}_j$ on a single image $\mathbf{x}_i$ with a mass-margin type y_i^{margin}, and comes with a fine-annotation mask $\mathbf{m}_i$, is defined as:

$$\sum_{j:\ \text{class}(\mathbf{p}_j)=y_i^{\text{margin}}} \left(\frac{\sum \left[(1-\mathbf{m}_i) \odot T_\tau \left(\text{PAM}_{i,j} \right) \right]}{\sum T_\tau \left(\text{PAM}_{i,j} \right)} \right), \qquad (1)$$

where T_τ is a threshold function that returns the top $(1-\tau)\times 100\%$ of the input values as 1 and the bottom $\tau \times 100\%$ as 0. The prototype activation map ($\text{PAM}_{i,j}$) shows where $\mathbf{p}_j$ is activated on $\mathbf{x}_i$. To evaluate activation precision for GradCAM (ref. [44]) and GradCAM++ (ref. [45]), we calculate as per equation (1) but replace $\text{PAM}_{i,j}$ with the normalized gradient map for the correct class; 95% confidence intervals were derived using non-parametric bootstrap resampling with 5,000 samples each equal to the size of the test set. Further information on activation precision can be found in the Supplementary Information.

Activation precision is a measure of interpretability, in the sense that the higher the activation precision, the better a prototype (or a set of prototypes) is at detecting medically relevant features for mass-margin classification.

Activation precision can be measured both at lesion-scale (that is, is the activation within the lesion area and not the added context window?) and at fine-scale (that is, is the activation on the specific part of the margin marked relevant by the radiologist?).

Mass margin classification. We compare the following models.

IAIA-BL. We used ProtoPNet with VGG-16 pre-trained on ImageNet as the base architecture trained for 50 epochs because model training converges between 40 and 50 epochs. The final

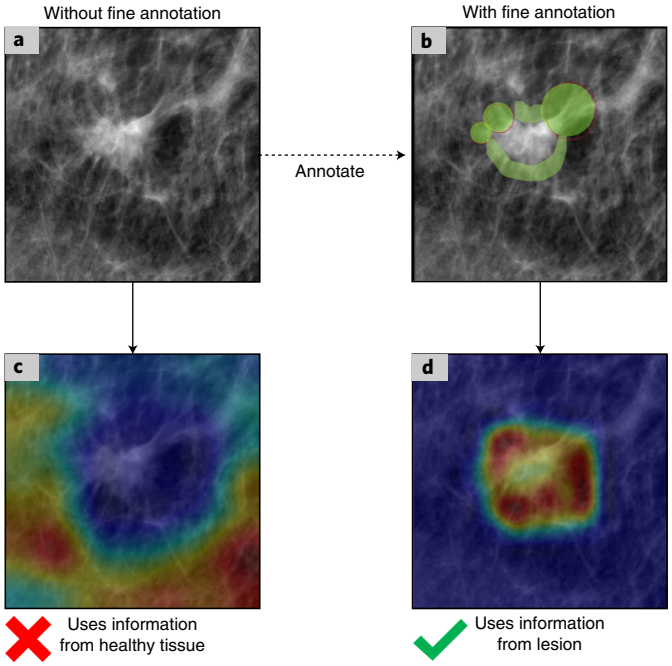


Fig. 2 | Fine-annotation regularization on model attention that penalizes the model for using confounding information. a, The lesion to be classified. **b**, The spicules of the lesion have been marked in green by a radiologist. **c**, Without fine annotation loss in training, the activation map highlights confounding information. **d**, Using fine annotation loss during training, the activation map highlights relevant information—areas that contain spicules. The attention is within the area marked by the radiologist (otherwise it would have been penalized by the fine-annotation loss function).

model is trained on the combination of the training set and validation set, and tested on a test set never before seen in training. See the Supplementary Information for hyperparameter tuning information. Our model can be fully trained on one P100 GPU in 50 h.

Original ProtoPNet. The original ProtoPNet (ref. [9]) architecture does not use fine annotation loss, and uses max pooling logic where IAIA-BL uses top-k average pooling logic. This change is equivalent to changing all uses of AVGPOOL to max in equations (6) and (9); and changing mink to min in equation (7).

VGG-16 with GradCAM and GradCAM++. We trained a VGG-16 (ref. [46]) model with added parameters to account for the larger number of parameters in our model. Pre-trained on ImageNet, it was trained for 250 epochs and the epoch with the highest test accuracy was selected for comparison. There is no native way to incorporate our fine annotation into VGG-16. VGG-16 provides no inherent interpretability or localization. Using the post-hoc GradCAM (ref. [44]) and GradCAM++ (ref. [45]) techniques we show localization information and calculate activation precision.

Results. Treating the radiologist annotations as the ground truth, results are reported in Table 1. Figure 3 shows ROC curves for all prediction tasks and all methods. Each mass margin prediction is explained as in Fig. 4. Extended Data Figs. 1–4 show sample explanations automatically generated by IAIA-BL.

IAIA-BL achieves an AUROC as good or better than the AUROCs of interpretable ProtoPNet and the analogous black-box model, VGG-16. Both ProtoPNet and VGG-16 show much lower activation precision than IAIA-BL. Both use information from image regions entirely outside the region that contains the lesion. The baseline models are not restricted from using confounding information, and thus do so freely. These models should not be used in practice for this reason. A visual comparison of activation maps (defined in ref. [9]) is shown in Extended Data Fig. 5.

For IAIA-BL without pruning, there is a 0.004 increase in AUROC for mass margin prediction. For the unpruned IAIA-BL model (not shown in the table, as it is almost identical to IAIA-BL), the lesion-scale activation precision of the learned prototypes is 0.93 (95% CI: 0.91, 0.96) and the fine-scale activation precision of the learned prototypes is 0.41 (95% CI: 0.39, 0.43).

Another measure of performance we calculated is the agreement in Cohen's κ between IAIA-BL and our human mass margin annotator on the test set. We found substantial agreement with a κ-value of 0.74 ($n=78$, 95% CI: 0.60, 0.86)[47], further broken down into circumscribed at 0.76 (95% CI: 0.58, 0.90), indistinct at 0.69 (95% CI: 0.51, 0.84) and spiculated at 0.78 (95% CI: 0.61, 0.93). For this task of characterizing the mass margin in mammography, our performances were higher than the inter-observer agreement between radiologists (for example, 0.61–0.65 in ref. [41], 0.58 in ref. [42] and 0.48 in ref. [43]).

Although there are many papers on computer vision with applications to mammography, few papers attempt to classify masses by margin type. We found only one study[48], which reports an accuracy of mass margin prediction at 70.6% and includes more margin types than we do, but their provided results are not separated into different margin classes and therefore we cannot directly compare with them. Moreover, that study used digitized mammography images from the DDSM database, further reducing comparability. Their model is not publicly available.

Malignancy prediction. IAIA-BL converts unnormalized mass margin scores ($\hat{\mathbf{y}}^{\text{margin}}$) to malignancy scores ($\hat{y}_i^{\text{mal}}$) with the following concise linear model learned in training stage B:

$$\hat{y}_i^{\text{mal}} = -16\,\hat{y}_i^{\text{circumscribed}} - 10\,\hat{y}_i^{\text{indistinct}} + 6\,\hat{y}_i^{\text{spiculated}}, \text{ with} \quad (2)$$

$$\text{Prob}(\text{malignancy}) = \sigma((\hat{y}_i^{\text{mal}} - 155)/100), \quad (3)$$

where $\sigma(t)$ is the logistic sigmoid function.

This model is consistent with medical knowledge: a high spiculated score results in a high probability of malignancy, whereas high circumscribed or indistinct margin scores indicate a benign lesion. Each mass margin score is explained in Fig. 4.

Although IAIA-BL is constrained to using only the results of the mass margin outputs to predict malignancy (rather than extra information that may be contained within the raw pixels of the image), IAIA-BL predicts mass malignancy with an AUROC of 0.84 ($n=75$, 95% CI: 0.74, 0.94). These results are interpretable in that they only use the mass margin scores to make their predictions.

We remark that the prediction of whether a mass has 5% or 95% probability of being malignant would not alter the clinical management, as all lesions with >2% probability of malignancy would be recommended to undergo breast biopsy. There are a variety of malignancy performance values reported in the literature, though not necessarily from the same population as ours, which means the results are not directly comparable. Some studies have reported better performance in predicting malignancy from BI-RADS features[49–51]. If our dataset were larger, and if we had non-imaging features such as patient age, it could potentially boost performance.

Table 1 | Mass-margin classification test results

	Model			
			VGG-16 (ref. [46])	VGG-16 (ref. [46])
	IAIA-BL	ProtoPNet (ref. [9])	with GradCAM (ref. [44])	with GradCAM++ (ref. [45])
Performance (AUROC)				
Mass margin classification	**0.951** [0.905, 0.996]	0.911[a] [0.848, 0.974]	0.947[a] [0.898, 0.996]	0.947[a] [0.898, 0.996]
Spiculated versus all	0.96[a] [0.90, 1.00]	**0.97** [0.93, 1.00]	0.95[a] [0.89, 1.00]	0.95[a] [0.89, 1.00]
Indistinct versus all	0.93[a] [0.88, 0.99]	0.87[a] [0.78, 0.94]	**0.94** [0.89, 0.99]	**0.94** [0.89, 0.99]
Circumscribed versus all	**0.97** [0.94, 1.00]	0.93[a] [0.87, 1.00]	0.95[a] [0.91, 1.00]	0.95[a] [0.91, 1.00]
Cohen's κ	**0.74** [0.60, 0.86]	0.64[a] [0.49, 0.78]	**0.74** [0.60, 0.87]	**0.74** [0.60, 0.87]
Interpretability				
Fine-scale act. prec.	**0.41** [0.39, 0.45]	0.24 [0.17, 0.31]	0.21 [0.05, 0.43][b]	0.24 [0.08, 0.45][b]
Lesion-scale act. prec.	**0.94** [0.92, 0.97]	0.51 [0.34, 0.68]	0.45 [0.37, 0.54][b]	0.53 [0.44, 0.61][b]

The first five rows measure prediction performance, whereas the lower two rows measure interpretability performance. The table shows that IAIA-BL's test AUROC performance with respect to all tasks is approximately as good as the best of the baselines. IAIA-BL's main advantage (interpretability) is shown in the bottom two rows of the table, where there is a drop in fine activation precision (act. prec.) for original ProtoPNet and VGG-16 as compared with IAIA-BL. VGG-16 has no inherent interpretability but post-hoc GradCAM and GradCAM++ provide localization information on which we measure activation precision. The best value is in bold. [a]Values not significantly below the best by Delong's test with $p<0.05$. [b]As this technique is post-hoc, there is no guarantee that the generated explanation matches the model's decision making.

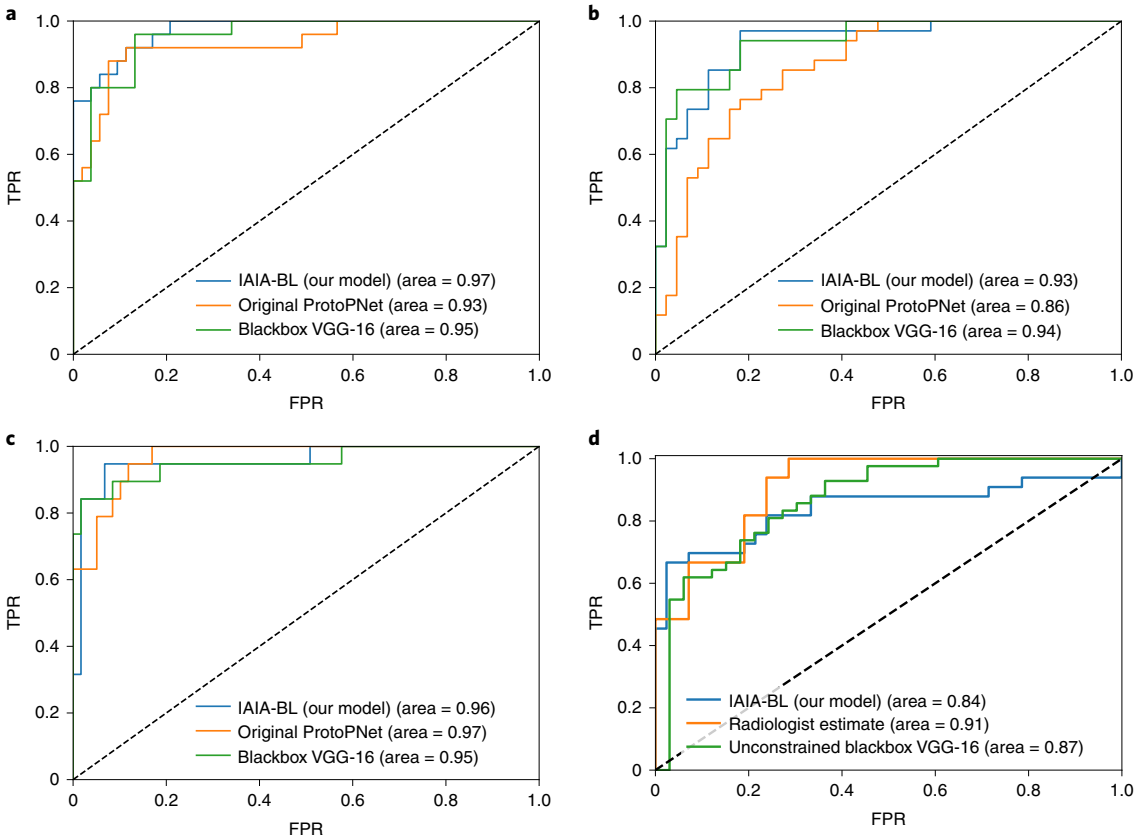


Fig. 3 | ROC curves of IAIA-BL compared with baselines. a, Circumscribed class versus other margin classes. **b**, Indistinct class versus other margin classes. **c**, Spiculated class versus other margin classes. **d**, Malignant versus benign lesions.

Radiologist estimate. During data collection, we asked radiologists to estimate the probability that the lesion will be malignant. There are several caveats for this estimate: radiologists do not perform this task in standard practice, instead they only provide a categorical recommendation for biopsy; the annotations were completed as part of a separate study that used consumer-grade monitors without the necessary specifications or calibrations of medical-grade displays. Nonetheless, these estimates represent the radiologist's best guess when given even more information than the model is provided. The radiologists predicted mass malignancy on the test set with AUROC of 0.91 ($n=75$, 95% CI: 0.85, 0.97). These radiologists are from Duke Hospital and thus represent an extremely high quality of care for patients. Using this as a reference standard, IAIA-BL is approximately 7% in AUROC away from the physicians.

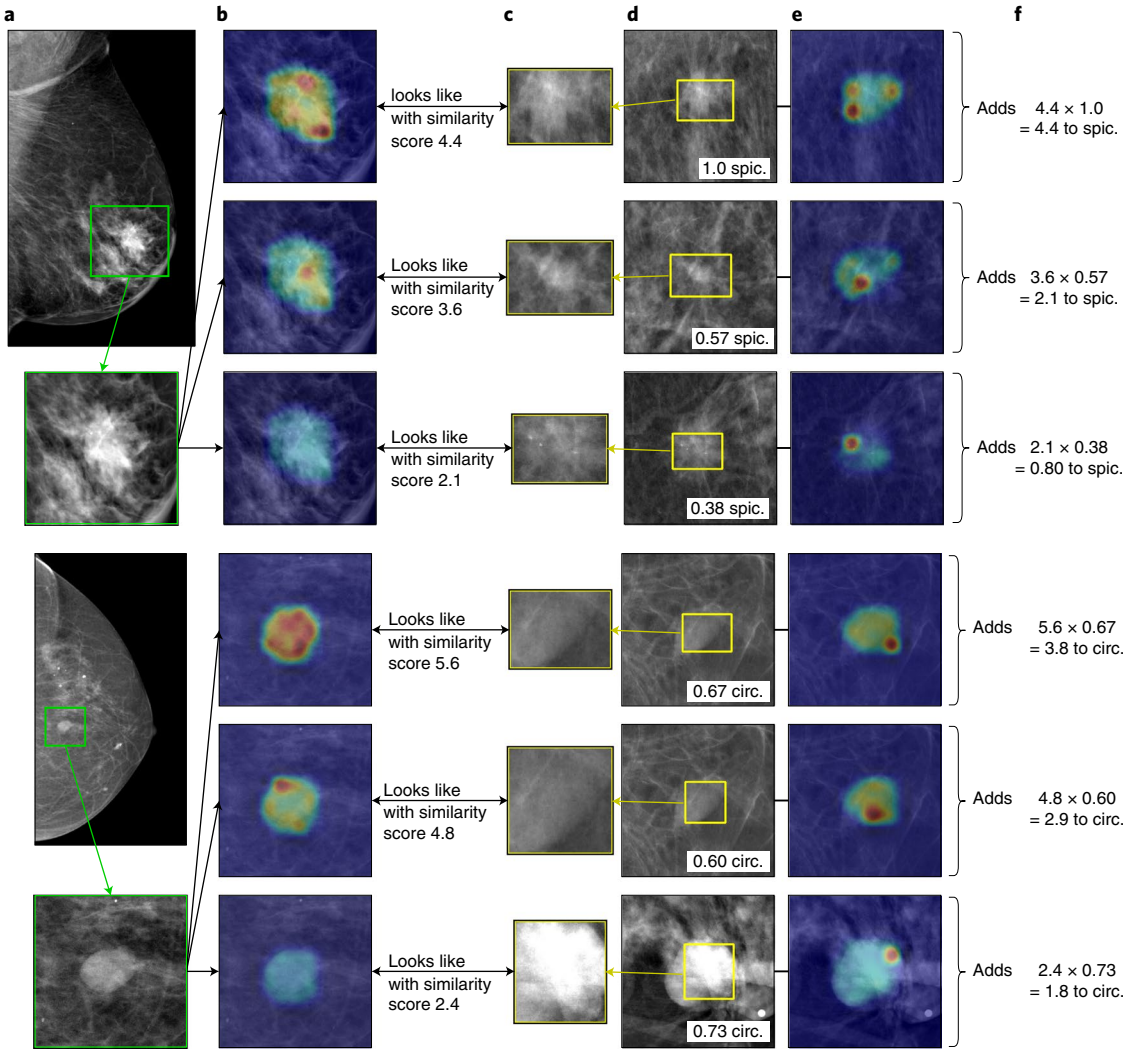


Fig. 4 | Case-based explanations generated by IAIA-BL. Top: This spiculated (spic.) lesion is correctly classified as spiculated. Bottom: This circumscribed (circ.) lesion is correctly classified as circumscribed. **a**, Test images. **b**, Activation of prototype on test images. **c**, Most relevant part of prototype. **d**, Learned prototypical lesion. **e**, Prototype self-activation. **f**, Contribution to class score.

Unrestricted end-to-end VGG-16. The uninterpretable VGG-16 baseline given the same image data, but not restricted to predicting on only mass margin results, achieves an AUROC of 0.87 ($n=75$, 95% CI: 0.82, 0.93). Again, it is possible that VGG-16 uses indirect information like breast density or confounding factors such as mammography equipment type.

We compared selected AI mammography techniques to IAIA-BL in Table 2. We cannot compare with papers focused on detection because our technique works on diagnosis of an already detected lesion[8,13].

Discussion

The high performance of uninterpretable models that seem to be leveraging mainly confounding information is a point of concern when incorporating models into clinical practice. Although a radiologist may not choose to view an explanation for every prediction, interpretable models still provide value over uninterpretable models. As we know that AI systems fail[4], we designed a system that can alert a radiologist to faulty reasoning at the time the prediction is made instead of only after the consequences of misprediction have been realized. The global interpretability (namely, the set of prototypes) allows the trained model to be fine-tuned by domain experts through pruning of prototypes that do not correspond to medically relevant features. The explanations provided can also be used for debugging a model and for retrospective analysis of model failures.

Our technique could be expanded with little change to include other BI-RADS features (for example, mass shape). The technique might be able to be expanded to microcalcification clusters—the other main type of breast lesions—but there are more categories of calcification morphologies and the different types of cluster distributions can translate into lesions with extreme differences in scale, which might pose interesting technical challenges. The underlying logic of the technique could be extended to digital breast tomosynthesis by representing a prototype as either a two-dimensional part of a reconstructed slice image, or as a three-dimensional portion of a tomosynthesis volume.

Future work with this model might include reader studies in which we measure any improvements in accuracy and radiologists report their trust in our system. Given the increased benefit of other AI assistance to less-experienced readers[52,53], it might be valuable to compare the benefit of this system with both sub-specialists and community radiologists who might be called on to do this work only occasionally. This work might help to extend the quality of care

Table 2 | Comparison of selected AI mammography techniques to IAIA-BL

	IAIA-BL	Ref. [13]	Ref. [48]	Ref. [56]
Inherently interpretable model (not post-hoc)	✓	✓	✓	
Provides global interpretability (on model)	✓			✓
Provides local interpretability (on each case)	✓	✓	✓	
Explanation is guaranteed to match model reasoning	✓			
Incorporate domain-specific terminology	✓		✓	✓
Provides similar prototypes for comparison	✓			
Can incorporate fine annotation	✓			
Can be trained on data with mixed labeling	✓	✓		

that patients receive at Duke (with highly trained Duke radiologists) to patients that do not have access to this level of care.

The fine annotation techniques we developed to reduce the use of confounding information can be extended to other computer vision applications. The fine annotation technique could also be used on datasets with known confounders to see how effectively it reduces (or reveals) use of the confounders in its classification decisions.

Conclusion

Our work shows that we are able to create interpretable mass margin prediction models with equal or higher performance to their uninterpretable counterparts. Using only a small dataset, we were able to provide an interpretable network that performs comparably with radiologists on mass margin classification and malignancy prediction. The gradient stabilization improvement to the ProtoPNet training can be added into any future use of its codebase.

Methods

Data augmentation. Given the small training set, we performed data augmentation such that each training image was randomly flipped, rotated and underwent random cropping with a crop size of 80% of the image's original size. Each class was augmented to have 5,000 images for the training set.

Framework. Models need to be collaborators in the medical decision-making process to be useful. In mammography, the initial clinical decision is expressed as a BI-RADS category of 1 to 5, corresponding to the recommendation of whether the patient needs a biopsy[7,54]. An inscrutable model predicting whether a lesion is malignant or benign is not useful as a decision aid, as a biopsy is recommended for every lesion with a greater than 2% chance of malignancy (BI-RADS 4 and 5). To alter clinical management, an interpretable model is needed to describe its reasoning process for why the patient should or should not receive a biopsy rather than provide an inscrutable prediction of malignancy.

Our AI approach includes an explicit reasoning system that resembles that of a practicing radiologist. Existing interpretability techniques for mammography include localization as in Fig. 1b, but there is no explanation of why an area is selected, what attributes of the region are used for classification or which parts of the training set these associations are learned from. In a non-medical image analogy, although localization may provide a good interpretation for whether or not an image contains a vase (perhaps by highlighting the vase), it does not provide a good interpretation for classification of the vase pattern as Roman versus Asian antiquity (highlighting the vase pattern provides no further insight). Many recently published AI mammography algorithms are still entirely uninterpretable, as in Fig. 1a (ref. [8]). Our method is designed to analyse lesions in a way that mimics how a radiologist would approach image interpretation (as in Fig. 1c), it first finds medically relevant features (in this case the mass margin) then uses those features to determine the possibility of malignancy. Such a model may be integrated into a clinical support system for classifying breast lesions, as it can point to mammogram regions that resemble prototypical signs of cancerous growth (for example, spiculated mass margin) and thereby assist doctors in making diagnoses.

Model architecture. Extended Data Fig. 6 gives an overview of our model architecture.

Given a region of interest $\mathbf{x}$ in a mammogram, our IAIA-BL model first extracts useful features $f(\mathbf{x})$ for mass-margin classification, using a series of convolutional layers f from a VGG-16 network[46] pre-trained on ImageNet. The output $f(\mathbf{x})$ is size $14\times14\times c$, where c is the number of channel; $l\in\{(1,1),\ldots,(1,14),(2,1),\ldots,(14,14)\}$ indexes each of the $1\times1\times c$ patches $f(\mathbf{x})_l$ across the spatial dimensions.

Following the convolutional layers f is prototype layer g. As in ProtoPNet[9], the prototype layer contains m prototypes $\mathbf{P}=\{\mathbf{p}_j\}_{j=1}^{m}$ learned from the training set. Each prototype is size $1\times1\times c$. As a prototype has the same c but a smaller spatial dimension than the convolutional feature maps $f(\mathbf{x})$, we can interpret the prototype as representing a prototypical activation pattern of its class and we can visualize the prototype as a patch of the training image it appears in.

Given convolutional feature maps $f(\mathbf{x})$, the prototype layer g calculates the similarity score s_j between $\mathbf{x}$ and each $\mathbf{p}_j$. It first computes the distance between $\mathbf{p}_j$ and each of the $l1\times1$ spatial patches of convolutional feature map $f(\mathbf{x})$ by: $d_{j,l}=\|\mathbf{p}_j-f(\mathbf{x})_l\|_2^2$, and converts distances to similarities:

$$s_{j,l}=\log\frac{d_{j,l}+1}{d_{j,l}+\epsilon}. \tag{4}$$

This provides a set of similarity scores $\{s_{j,l}\}_{l=(1,1)}^{(14,14)}$ that can be arranged spatially into a similarity map $[s_{j,l}]_{l=(1,1)}^{(14,14)}$ that compares the input image and each $\mathbf{p}_j$. The overall s_j is calculated using top-k average pooling (as in ref. [55]):

$$s_j=\operatorname{avg}\left(\operatorname{top}_k\left(\{s_{j,l}\}_{l=(1,1)}^{(14,14)}\right)\right). \tag{5}$$

Conceptually, this means that if $\mathbf{x}$ has spicules along the mass margin, its convolutional feature maps $f(\mathbf{x})$ will have patches $f(\mathbf{x})_l$ that represent the spicules from the input image. These patches will be close (in ℓ_2 distance in the latent space) to one or more $\mathbf{p}_j$ that represent spicules on the mass margin. Consequently, $s_{j,l}$ will be large between those spiculated prototypes and patches.

In IAIA-BL we initialize the model with $m=15$. We prune duplicate prototypes and the final IAIA-BL model presented has four prototypes for a circumscribed mass margin, three for an indistinct mass margin and four for a spiculated mass margin. We set c to 512 in our experiments.

IAIA-BL ends with two fully connected layers. The first fully connected layer h_1 multiplies the vector of similarity scores $[s_1,\ldots,s_m]$ by a weight matrix to produce three output scores: $\hat{y}^{\text{circumscribed}}$, $\hat{y}^{\text{indistinct}}$ and $\hat{y}^{\text{spiculated}}$, one for each margin type. These are (afterwards) normalized using a softmax function to generate the probabilities that the mass margin in the input image belongs to each of the three mass-margin types. The second fully connected layer h_2 then combines the vector of (unnormalized) mass-margin scores $\hat{\mathbf{y}}^{\text{margin}}=[\hat{y}^{\text{circumscribed}},\hat{y}^{\text{indistinct}},\hat{y}^{\text{spiculated}}]$ into a final score of malignancy $\hat{y}^{\text{mal}}$, which is passed into a logistic sigmoid function to produce a probability that the input image has a malignant breast cancer.

Fine annotations. When starting to build an interpretable model for breast lesion classification, we naively applied the case-based reasoning ProtoPNet architecture to medical images. Although the model seemed to be learning medically relevant features due to its high validation accuracy, the model made predictions using regions of the image that did not correspond to the medical information; in other

words, the model used confounding information rather than medically relevant information. This is consistent with observations made by other groups of the dangers of confounding in medical imaging[25]. For non-medical image classification tasks, a typical approach might be to increase the size of the training set; however, as discussed above, one major barrier to implementation of AI in the medical field is the limited availability of annotated data[6].

To extract more information from our dataset, we collected a small set of pixel-level (fine) annotations from our radiology team, which permitted better generalization using a smaller number of images; that is, fine annotation on only 30 images of our limited data set (1,136 annotated mammographic images from 484 patients with lesions) enabled high-quality reasoning and prediction. This novel approach can reduce the confounding in deep learning by leveraging both relatively abundant coarsely annotated data and a small amount of finely annotated data. Most fine-grained classification algorithms either assume the availability of fine-grained part labels for all input data and use them, or those fine-grained labels are completely ignored. Our approach provides a middle ground by using both data with and without fine-grained annotations, which takes full advantage of the information available. This approach is also practical in the sense that for real-world problems, annotated data are relatively less abundant and more expensive to obtain.

We designed a new training paradigm that incorporates this additional expert annotation information on a subset of the existing patient cases. A radiologist (F.R.S) annotated the area of a lesion image that indicates the mass margin for that lesion as in Fig. 2b, with the most prominent and defining features marked by circles and the rest of the lesion margin highlighted by simple lines. The model incorporates the radiologist-supplied fine annotations by regularizing the activation of the prototypes over the image. It penalizes a prototype for activating anywhere on an image not of its class, or for activating outside the region of the image marked relevant by the radiologist. Figure 2c shows an attention map that highlights confounding information and would be heavily penalized. Figure 2d shows an attention map that highlights relevant information. By directing the network to the most relevant parts of the image, we set a strong prior on the network for where the useful information is centered in the image. As these annotations are expensive to obtain, we designed the method to be able to use a small number of these finely annotated cases and a larger number of less expensive coarsely annotated cases. We include a training loss term in the objective as described in the methods. When using this, our performance and explanation both improve.

Model training. We represent the dataset of n $\mathbf{x}_i$, with y_i^{margin} and y_i^{mal} labels, as $D = \{(\mathbf{x}_i, y_i^{\text{margin}}, y_i^{\text{mal}})\}_{i=1}^{n}$. A thirty-image subset $D' \subseteq D$ comes with the radiologist's (fine) annotations of where medically relevant information is in that training image. For a training instance $(\mathbf{x}_i, y_i^{\text{margin}}, y_i^{\text{mal}}) \in D'$, we define $\mathbf{m}_i$ such that $\mathbf{m}_i$ takes the value 0 at those pixels that are marked as 'relevant to mass margin identification' and 1 at other pixels. Each $\mathbf{m}_i$ has the same spatial dimensions (height and width) as $\mathbf{x}_i$.

The training of IAIA-BL has four stages: (A1) training of the convolutional layers f and the prototype layer g; (A2) projection of prototypes; (A3) training of the first fully connected layer h_1 for predicting mass-margin types; and (B) training of the second fully connected layer h_2 for predicting malignancy probability. Stages A1, A2 and A3 are repeated until the training loss for predicting mass-margin types converges, after which we move to Stage B.

Stage A1. In the first training stage we aim to learn meaningful convolutional features. In particular, we want convolutional features that represent a particular mass-margin type to be clustered around a prototype of that particular mass-margin type, and to be far away from a prototype of other mass-margin types. As in Chen and colleagues[9], we jointly optimize the parameters $\boldsymbol{\theta}_f$ of f, and $\mathbf{p}_1, \ldots, \mathbf{p}_m$ of g, while keeping the two fully connected layers h_1 and h_2 fixed. We minimize the following training loss:

$$\min_{\boldsymbol{\theta}_f, \mathbf{p}_1, \ldots, \mathbf{p}_m} \text{CrsEnt} + \lambda_c \text{Clst} + \lambda_s \text{Sep} + \lambda_f \text{Fine}, \tag{6}$$

where the cross-entropy term (CrsEnt) penalizes misclassification of mass-margin types on the training data; it also ensures that the learned convolutional features and the learned prototypes are relevant for predicting mass-margin types.

Differing from Chen and colleagues[9] by the use of mink instead of min, the cluster (Clst) and separation (Sep) costs are defined by:

$$\begin{aligned} \text{Clst} &= \tfrac{1}{n} \sum_{i=1}^{n} \min_{j:\ \text{class}\,(\mathbf{p}_j) = y_i^{\text{margin}}} (\gamma), \\ \text{Sep} &= -\tfrac{1}{n} \sum_{i=1}^{n} \min_{j:\ \text{class}\,(\mathbf{p}_j) \neq y_i^{\text{margin}}} (\gamma), \ \text{with} \\ \gamma &= \tfrac{1}{k} \sum \text{mink}_{\mathbf{z} \in \text{patches}(f(\mathbf{x}_i))} \left(\| \mathbf{z} - \mathbf{p}_j \|_2^2 \right) \end{aligned} \tag{7}$$

where mink gives the k smallest squared distances. Empirically, we found that IAIA-BL trained with the relaxed cluster and separation costs outperforms the one trained with the original (that is, $k = 1$) cluster and separation costs of Chen and co-workers[9] on the task of margin classification, possibly because the relaxed cluster and separation costs (along with the top-k average pooling) allow the gradient of the loss function to back-propagate through k convolutional patches—instead of just one patch—during training. Consequently, the gradient will be less sensitive and more robust to changes in the location of the most activated convolutional patch by each prototype.

The fine-annotation loss, which is new to this paper, penalizes prototype activations on medically irrelevant regions of radiologist-annotated training mammograms (see Fig. 2). The fine-annotation loss is defined by:

$$\text{Fine} = \sum_{i \in D'} \left(\sum_{j:\ \text{class}\,(\mathbf{p}_j) = y_i^{\text{margin}}} \| \mathbf{m}_i \odot \text{PAM}_{i,j} \|_2 + \sum_{j:\ \text{class}\,(\mathbf{p}_j) \neq y_i^{\text{margin}}} \| \text{PAM}_{i,j} \|_2 \right) \tag{8}$$

where prototype activation map $\text{PAM}_{i,j}$ for $\mathbf{p}_j$ over $\mathbf{x}_i$ is computed by bilinearly upsampling the similarity map $[s_{j,l}]_{l=(1,1)}^{(14,14)}$ to yield the the same dimensions (height and width) as the fine-annotation mask. This promotes the learning of prototypes that stay away from any features that could appear in classes that are not the prototypes's designated classes, so that the prototypes of a particular class represent distinguishing features of that class.

To incorporate the training data with fine annotations into model training, we construct batches with 75 training examples from D with lesion-scale annotations and ten training examples from D' with fine-scale annotations. The fine-annotation loss on a lesion-scale annotation penalizes activation outside of the area marked as the lesion, whereas the fine-annotation loss on a finely annotated image penalizes activation outside of the region 'relevant to the mass margin class', as marked by the radiologist.

The prototype layer was initialized randomly using the uniform distribution over a unit hypercube (as the convolutional features from the last convolutional layer all lie between 0 and 1). In our experiments, $\lambda_c = 0.8$, $\lambda_s = 0.08$ and $\lambda_f = 0.001$.

Stage A2. As in work of Chen and colleagues[9], we project $\mathbf{p}_j$ onto the nearest convolutional feature patch from the training set D, of the same class as $\mathbf{p}_j$.

Stage A3. In this stage, we fine-tune the first fully connected layer h_1 to further increase the accuracy in predicting mass-margin types. We fix $\boldsymbol{\theta}_f$ and $\mathbf{p}_1, \ldots, \mathbf{p}_m$, and minimize the following training objective with respect to $\boldsymbol{\theta}_{h_1}$ of the first fully connected layer h_1:

$$\min_{\boldsymbol{\theta}_{h_1}} \frac{1}{n} \sum_{i=1}^{n} \text{CrsEnt}(h_1 \circ g \circ f(\mathbf{x}_i), y_i^{\text{margin}}). \tag{9}$$

The first time we enter stage A3, we initialize connections in fully connected layer h_1 to a value of 1 for prototypes that are positive for that mass margin, −1 otherwise.

Stage B. In this stage, we train the second fully connected layer h_2 for predicting malignancy probability, using a logistic regression model whose input is the (unnormalized) mass-margin scores produced by the first fully connect layer h_1, and whose output is the malignancy score. To prevent the malignancy information from biasing the mass margin classification, we train the model in a modular style and it is not trained completely end-to-end in any stage, that is, there is no return to Stage A from Stage B.

Reporting Summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The imaging data are not publicly available because they contain confidential information that may compromise patient privacy as well as the ethical or regulatory policies of our institution. Data will be made available on reasonable request, for non-commercial research purposes, to those who contact J.L. (joseph.lo@duke.edu). Data usage agreements may be required. Source Data are provided with this paper.

Code availability

Code is available on GitHub at https://github.com/alinajadebarnett/iaiabl. Two licenses are offered: an MIT license for non-commercial use and a custom license. The doi for the initial code release is https://doi.org/10.5281/zenodo.5565592.

Received: 12 March 2021; Accepted: 24 October 2021;
Published online: 15 December 2021

References

1. Kochanek, K. D., Xu, J. & Arias, E. *Mortality In the United States, 2019* Techical Report 395 (NCHS, 2020); https://www.cdc.gov/nchs/products/databriefs/db395.htm
2. Badgeley, M. A. et al. Deep learning predicts hip fracture using confounding patient and healthcare variables. *NPJ Digit. Med.* **2**, 1–10 (2019).
3. Winkler, J. K. et al. Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition. *JAMA Dermatol.* **155**, 1135–1141 (2019).
4. Zech, J. R. et al. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS Med.* https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1002683 (2018).
5. Edwards, B. FDA Guidance on clinical decision support: peering inside the black box of algorithmic intelligence. *ChilmarkResearch* https://www.chilmarkresearch.com/fda-guidance-clinical-decision-support/ (2017).
6. Soffer, S. et al. Convolutional neural networks for radiologic images: a radiologist's guide. *Radiology* **290**, 590–606 (2019).
7. Sickles, E et al. in. *ACR BI-RADS Atlas, Breast Imaging Reporting and Data System.* 5th edn, (American College of Radiology, 2013).
8. McKinney, S. M. et al. International evaluation of an ai system for breast cancer screening. *Nature* **577**, 89–94 (2020).
9. Chen, C. et al. This looks like that: deep learning for interpretable image recognition. In *Advances in Neural Information Processing Systems 32* 8930–8941 (NeurIPS, 2019).
10. Lehman, C. D. et al. Diagnostic accuracy of digital screening mammography with and without computer-aided detection. *JAMA Internal Med.* **175**, 1828–1837 (2015).
11. Salim, M. et al. External evaluation of 3 commercial artificial intelligence algorithms for independent assessment of screening mammograms. *JAMA Oncol.* **6**, 1581–1588 (2020).
12. Schaffter, T. et al. Evaluation of combined artificial intelligence and radiologist assessment to interpret screening mammograms. *JAMA Network Open* **3**, e200265– (2020).
13. Wu, N. et al. Deep neural networks improve radiologists' performance in breast cancer screening. *IEEE Trans. Med. Imaging* **39**, 1184–1194 (2019).
14. Kim, H.-E. et al. Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study. *The Lancet Digital Health* **2**, e138–e148 (2020).
15. Giger, M. L., Chan, H.-P. & Boone, J. Anniversary paper: history and status of CAD and quantitative image analysis: the role of medical physics and AAPM. *Med. Phys.* **35**, 5799–5820 (2008).
16. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **1**, 206–215 (2019).
17. Adebayo, J. et al. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems* 9505–9515 (NeurIPS, 2018).
18. Arun, N. et al. Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging. *Radiology: Artificial Intelligence* **3** (2021).
19. Wu, T. & Song, X. Towards interpretable object detection by unfolding latent structures. In *Proc. IEEE International Conference on Computer Vision* 6033–6043 (IEEE, 2019).
20. Chen, Z., Bei, Y. & Rudin, C. Concept whitening for interpretable image recognition. *Nat. Mach. Intell.* **2**, 772–782 (2020).
21. Demigha, S. & Prat, N. A case-based training system in radiology-senology. In *Proc. 2004 International Conference on Information and Communication Technologies: From Theory to Applications, 2004* 41–42 (IEEE, 2004).
22. Macura, R. T. & Macura, K. J. Macrad: Radiology image resource with a case-based retrieval system. In *International Conference on Case-Based Reasoning* 43–54 (Springer, 1995).
23. Floyd Jr, C. E., Lo, J. Y. & Tourassi, G. D. Case-based reasoning computer algorithm that uses mammographic findings for breast biopsy decisions. *Am. J. Roentgenol.* **175**, 1347–1352 (2000).
24. Kobashi, S., Kondo, K. & Hata, Y. Computer-aided diagnosis of intracranial aneurysms in MRA images with case-based reasoning. *IEICE Trans. Inform. Syst.* **89**, 340–350 (2006).
25. Wang, H., Wu, Z. & Xing, E. P. Removing confounding factors associated weights in deep neural networks improves the prediction accuracy for healthcare applications. *Pac. Symp. Biocomput.* **24**, 54–65 (2019).
26. Hu, S., Ma, Y., Liu, X., Wei, Y. & Bai, S. Stratified rule-aware network for abstract visual reasoning. In *AAAIConference on Artificial Intelligence (AAAI)* (2021).
27. Dundar, A. & Garcia-Dorado, I. Context augmentation for convolutional neural networks. Preprint at https://arxiv.org/abs/1712.01653 (2017).
28. Xiao, K., Engstrom, L., Ilyas, A. & Madry, A. Noise or signal: The role of image backgrounds in object recognition. In *International Conference on Learning Representations* (2020).
29. Luo, J., Tang, J., Tjahjadi, T. & Xiao, X. Robust arbitrary view gait recognition based on parametric 3D human body reconstruction and virtual posture synthesis. *Pattern Recognition* **60**, 361–377 (2016).
30. Charalambous, C. & Bharath, A. A data augmentation methodology for training machine/deep learning gait recognition algorithms. In *Proc. British Machine Vision Conference (BMVC)* (eds Richard, C. et al.) 110.1–110.12 (BMVA, 2016).
31. Tang, R., Du, M., Li, Y., Liu, Z. & Hu, X. Mitigating gender bias in captioning systems. In *Proc. Web Conference 2021*, 633–645 (2021).
32. Zhao, Q., Adeli, E. & Pohl, K. M. Training confounder-free deep learning models for medical applications. *Nat. Commun.* **11**, 1–9 (2020).
33. Schramowski, P. et al. Making deep neural networks right for the right scientific reasons by interacting with their explanations. *Nat. Mach. Intell.* **2**, 476–486 (2020).
34. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A. & Torralba, A. Learning deep features for discriminative localization. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 2921–2929 (IEEE, 2016).
35. Zheng, H., Fu, J., Mei, T. & Luo, J. Learning multi-attention convolutional neural network for fine-grained image recognition. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 5209–5217 (IEEE, 2017).
36. Fu, J., Zheng, H. & Mei, T. Look Closer to See Better: Recurrent Attention Convolutional Neural Network for Fine-grained Image Recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4438–4446 (IEEE, 2017).
37. DeLong, E. R., DeLong, D. M. & Clarke-Pearson, D. L. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 837–845 (1988).
38. Sun, X. & Xu, W. Fast implementation of delong's algorithm for comparing the areas under correlated receiver operating characteristic curves. *IEEE Signal Process. Lett.* **21**, 1389–1393 (2014).
39. Park, C. S. et al. Observer agreement using the ACR breast imaging reporting and data system (BI-RADS)-ultrasound, (2003). *Korean J. Radiol.* **8**, 397 (2007).
40. Abdullah, N., Mesurolle, B., El-Khoury, M. & Kao, E. Breast imaging reporting and data system lexicon for us: interobserver agreement for assessment of breast masses. *Radiology* **252**, 665–672 (2009).
41. Baker, J. A., Kornguth, P. J. & Floyd Jr, C. Breast imaging reporting and data system standardized mammography lexicon: Observer variability in lesion description. *AJR Am. J. Roentgenol.* **166**, 773–778 (1996).
42. Rawashdeh, M., Lewis, S., Zaitoun, M. & Brennan, P. Breast lesion shape and margin evaluation: Bi-rads based metrics understate radiologists' actual levels of agreement. *Comput. Biol. Med.* **96**, 294 – 298 (2018).
43. Lazarus, E., Mainiero, M. B., Schepps, B., Koelliker, S. L. & Livingston, L. S. Bi-rads lexicon for us and mammography: interobserver variability and positive predictive value. *Radiology* **239**, 385–391 (2006).
44. Selvaraju, R. R. et al. Grad-CAM: visual explanations from deep networks via gradient-based localization. In *The IEEE International Conference on Computer Vision (ICCV)* (IEEE, 2017).
45. Chattopadhay, A., Sarkar, A., Howlader, P. & Balasubramanian, V. N. Grad-CAM++: generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)* 839–847 (IEEE, 2018).
46. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. In *Proc. 3rd International Conference on Learning Representations (ICLR)* (2015).
47. Landis, J. R. & Koch, G. G. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics* **33**, 363–374 (1977).
48. Kim, S. T., Lee, H., Kim, H. G. & Ro, Y. M. ICADx: interpretable computer aided diagnosis of breast masses. In *Medical Imaging 2018: Computer-Aided Diagnosis* Vol. 10575, 1057522 (International Society for Optics and Photonics, 2018).
49. Elter, M., Schulz-Wendtland, R. & Wittenberg, T. The prediction of breast cancer biopsy outcomes using two cad approaches that both emphasize an intelligible decision process. *Med. Phys.* **34**, 4164–4172 (2007).
50. Benndorf, M., Burnside, E. S., Herda, C., Langer, M. & Kotter, E. External validation of a publicly available computer assisted diagnostic tool for mammographic mass lesions with two high prevalence research datasets. *Med. Phys.* **42**, 4987–4996 (2015).
51. Burnside, E. S. et al. Probabilistic computer model developed from clinical data in national mammography database format to classify mammographic findings. *Radiology* **251**, 663–672 (2009).
52. Park, H. J. et al. A computer-aided diagnosis system using artificial intelligence for the diagnosis and characterization of breast masses on ultrasound: added value for the inexperienced breast radiologist. *Medicine* **98**, e14146 (2019).
53. Shimauchi, A. et al. Evaluation of clinical breast MR imaging performed with prototype computer-aided diagnosis breast MR imaging workstation: reader study. *Radiology* **258**, 696–704 (2011).
54. Orel, S. G., Kay, N., Reynolds, C. & Sullivan, D. C. Bi-rads categorization as a predictor of malignancy. *Radiology* **211**, 845–850 (1999).
55. Kalchbrenner, N., Grefenstette, E. & Blunsom, P. A convolutional neural network for modelling sentences. In *Proc. 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* 655–665 (2014).

56. Wu, J. et al. Deepminer: Discovering interpretable representations for mammogram classification and explanation. *Harvard Data Science Review* **3** (2021).

Acknowledgements

We would like to acknowledge breast radiologists M. Taylor-Cho, L. Grimm, C. Kim and S. Yoon, who annotated the dataset used in this paper. This study was supported in part by NIH/NCI U01-CA214183 and U2C-CA233254 (J.L.). This study was supported in part by MIT Lincoln Laboratory (C.R.), Duke TRIPODS CCF-1934964 (C.R.) and the Duke Incubation Fund (A.J.B.).

Author contributions

A.J.B., F.S., D.T., C.C., J.L. and C.R. conceived the idea and developed the model. D.T., A.J.B. and C.C. wrote and reviewed the code. Y.R., A.J.B., F.S. and J.L. performed data collection, and Y.R., D.T. and A.J.B. preprocessed it.

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s42256-021-00423-x.

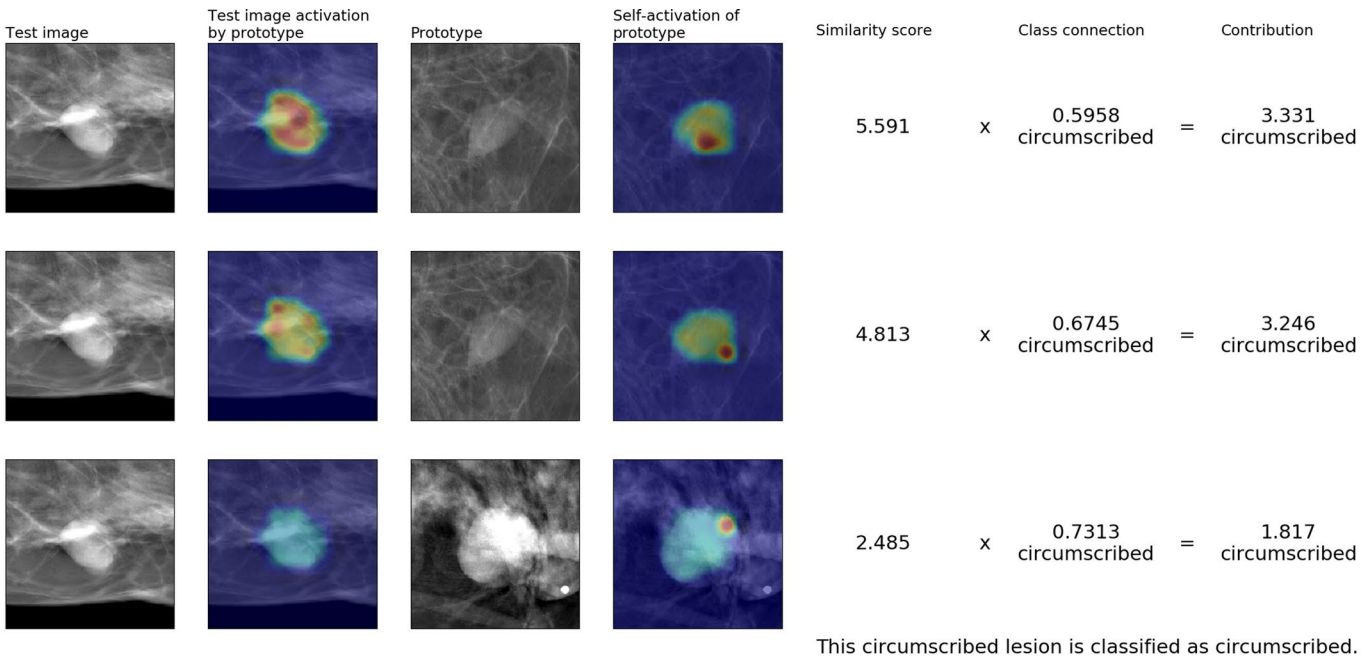
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42256-021-00423-x.

Correspondence and requests for materials should be addressed to Alina Jade Barnett.

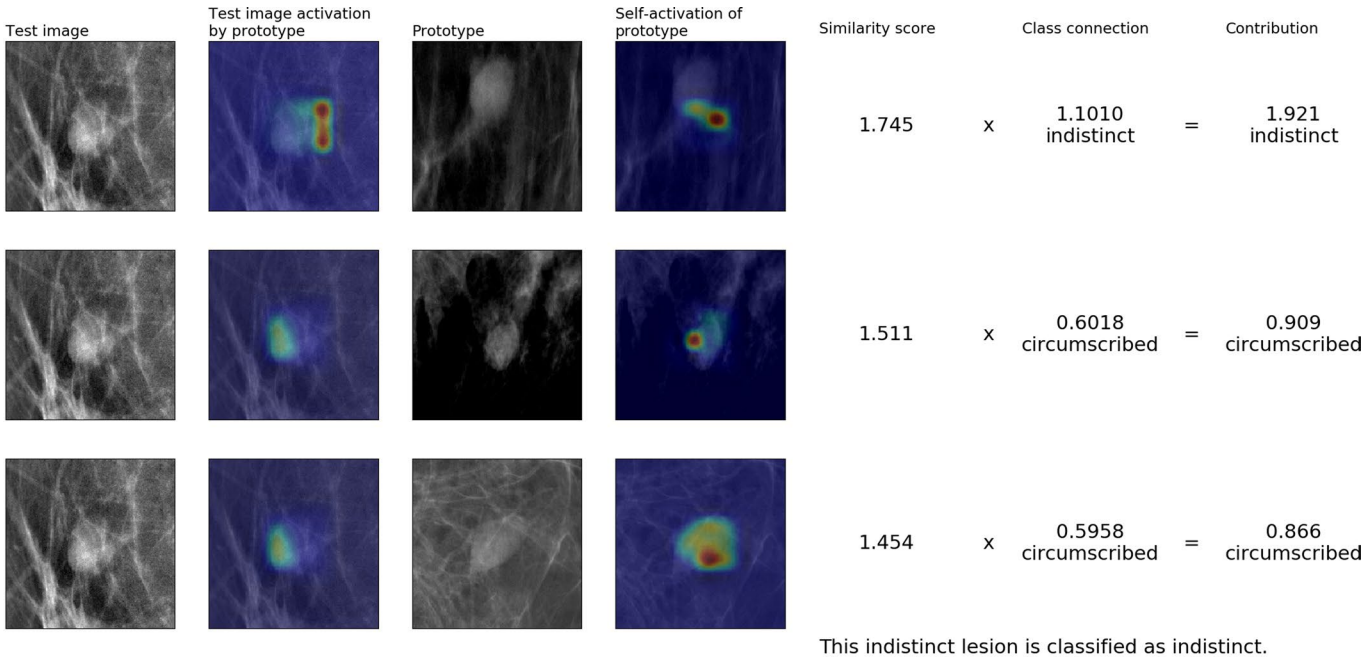
Peer review information *Nature Machine Intelligence* thanks Fredrik Strand and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

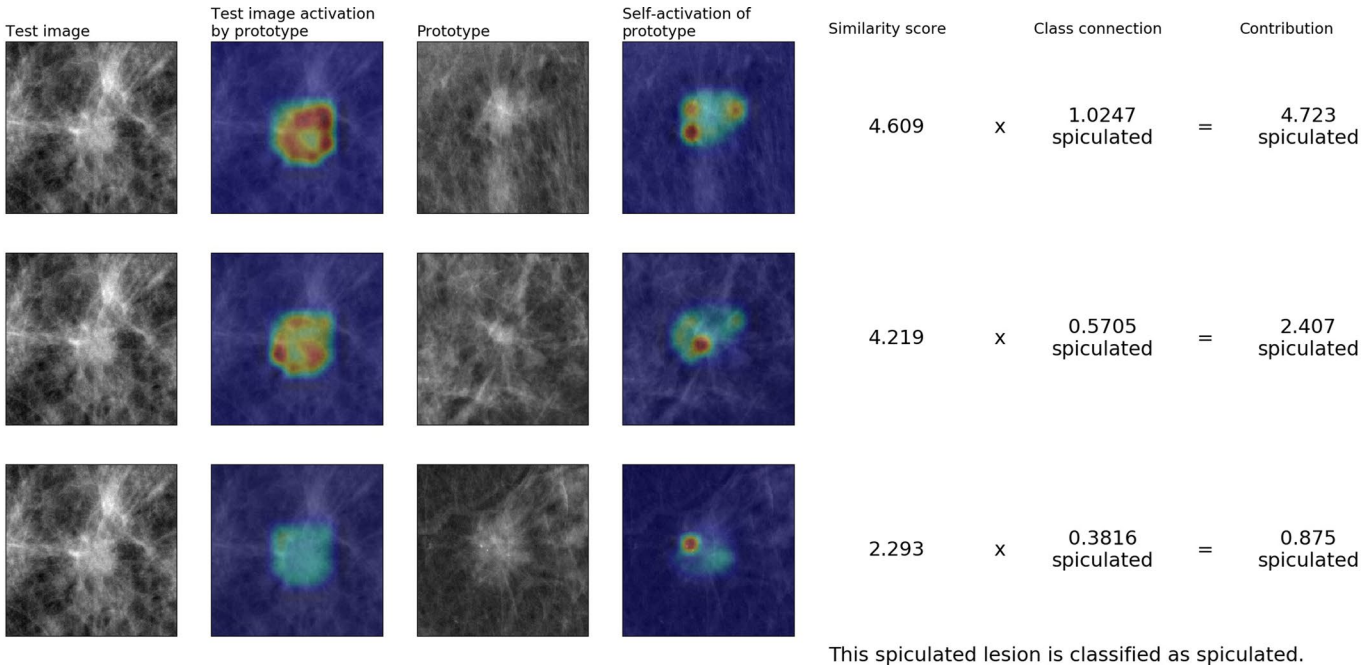
Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



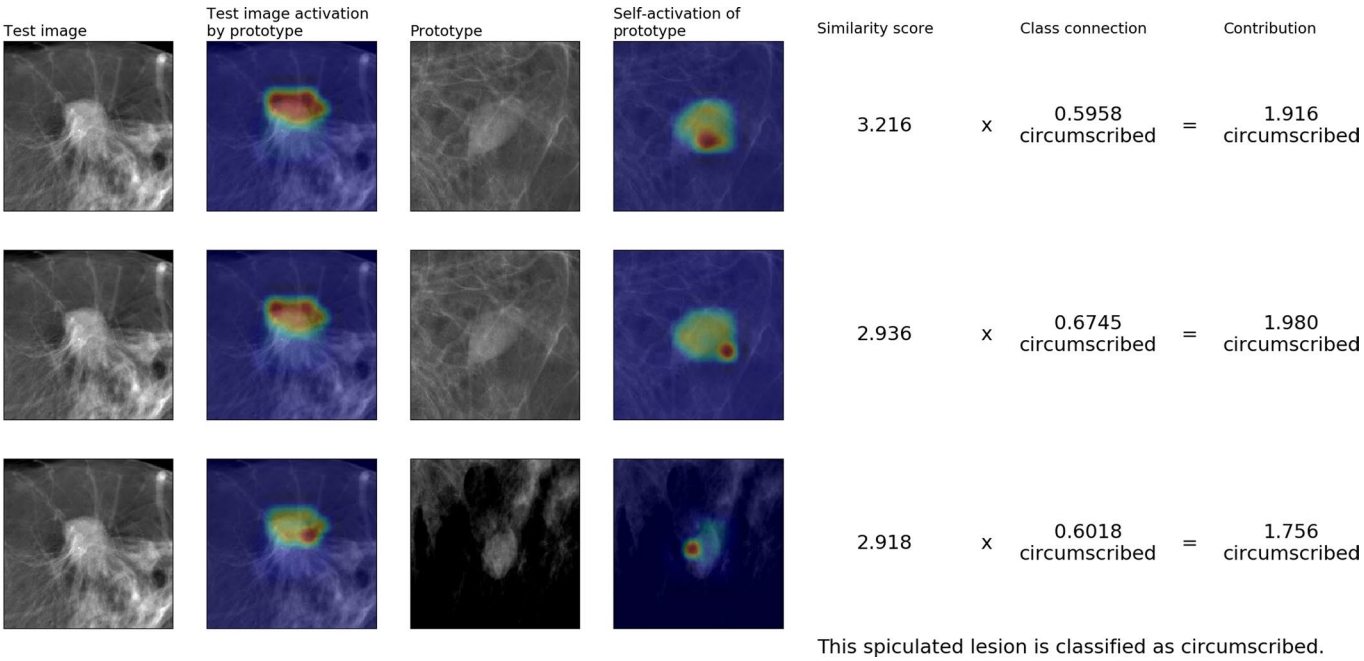
Extended Data Fig. 1 | An automatically generated explanation of mass margin classification for a circumscribed lesion. This circumscribed lesion is correctly identified as circumscribed. The first two most activated prototypes are drawn from the same image, but are associated with different regions of that image.



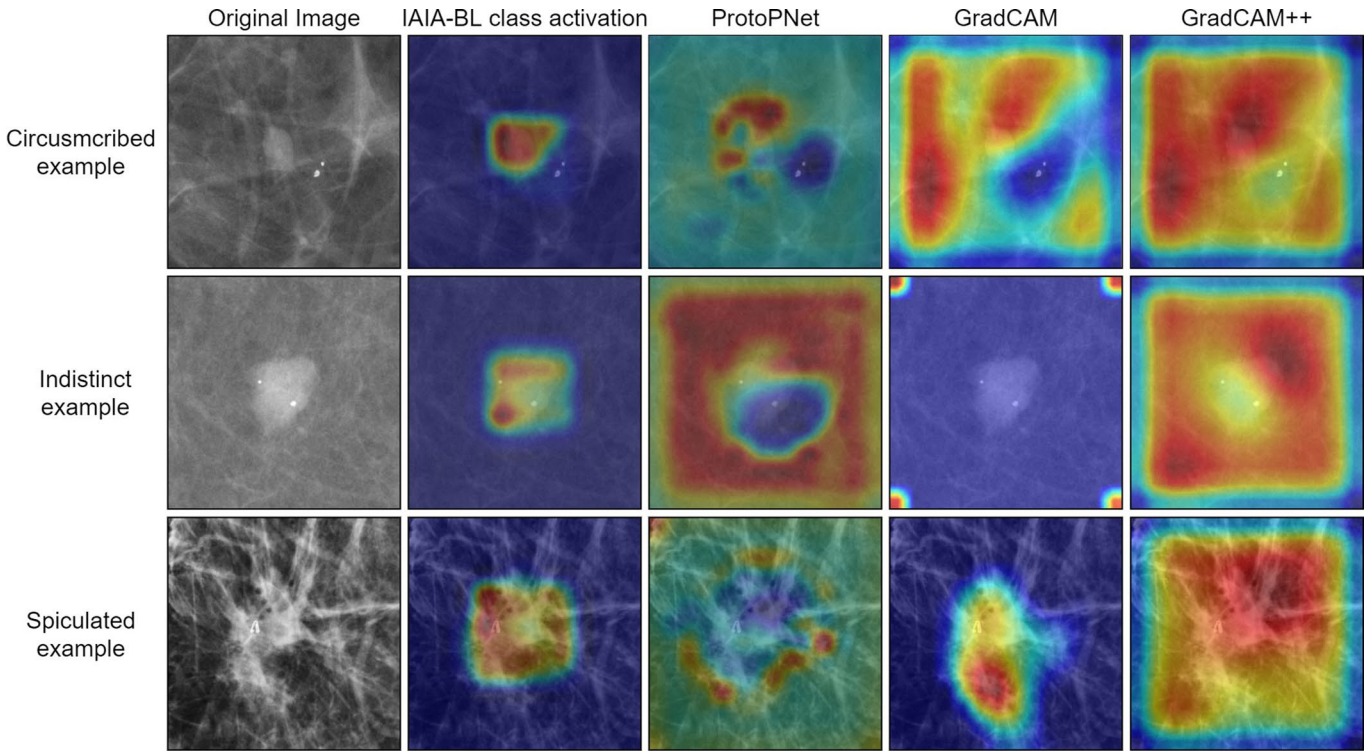
Extended Data Fig. 2 | An automatically generated explanation of mass margin classification for an indistinct lesion. This indistinct lesion is correctly identified as indistinct. The indistinct portion of the lesion margin (right side) activates the indistinct prototype and the circumscribed portion of the lesion margin (left side) activates the circumscribed prototypes.



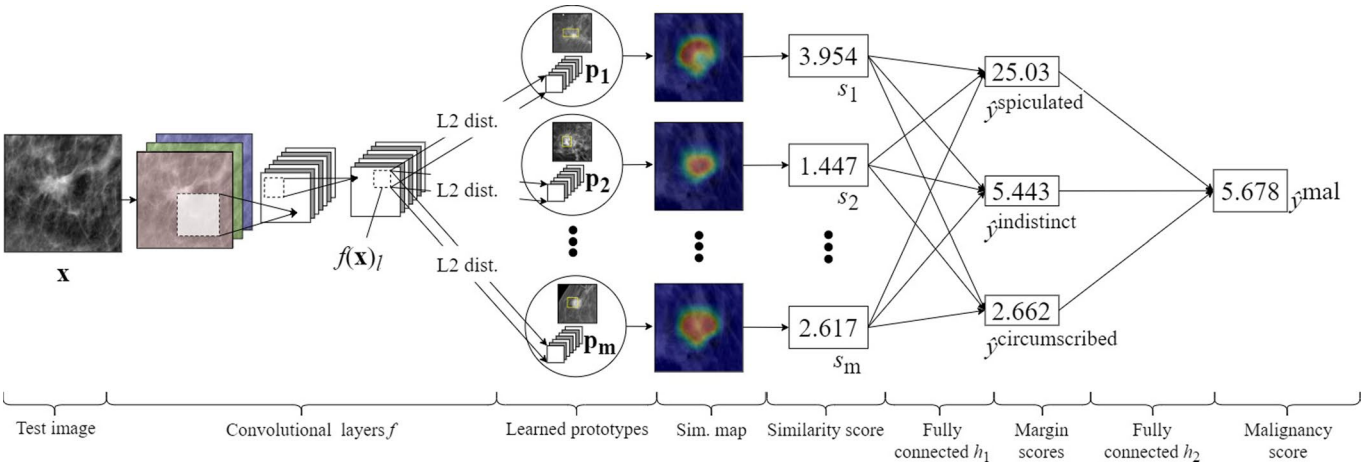
Extended Data Fig. 3 | An automatically generated explanation of mass margin classification for a spiculated lesion. This spiculated lesion is correctly identified as spiculated.



Extended Data Fig. 4 | An automatically generated explanation of mass margin classification for an incorrectly classified lesion. This spiculated lesion is incorrectly identified as circumscribed. The explanation highlights only the circumscribed portion of the mass margin (top), but does not detect the spiculated portion (bottom).



Extended Data Fig. 5 | A comparison of explanations. We compare explanations from two common saliency methods (GradCAM [44] and GradCAM++ [45]) to a class activation visualization derived from our method. The explanations from IAIA-BL are more likely to highlight the lesion and less likely to highlight the surrounding healthy tissue. This is shown quantitatively by the activation precision metric. The single image visualization is a dramatic simplification of the full explanation that is generated by IAIA-BL. The IAIA-BL and ProtoPNet class activation visualizations shown in this figure are generated by taking the average of prototype activation maps for all prototypes of the correct class.



Extended Data Fig. 6 | The architecture of the IAIA-BL prototype network. Test image **x** feeds into convolutional layers *f*. Each patch of $f(\mathbf{x})_l$ is compared to each learned prototype $\mathbf{p}_j$ by calculating the squared distance between the patch and the prototype. The similarity map shows the closest (most 'activated,' that is, smallest L^2 distance) patches in red and the furthest patches in blue, overlaid on the test image. Similarity score $\mathbf{s}_j$ is calculated from the corresponding similarity map. The similarity scores *s* feed into fully connected layer h_1, outputting margin logits $\hat{\mathbf{y}}^{\text{margin}}$. Margin logits $\hat{\mathbf{y}}^{\text{min}}$ feed into fully connected layer h_2, outputting malignancy logit y^{mal}.

nature research

Corresponding author(s): Alina Jade Barnett

Last updated by author(s): Oct 13, 2021

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Data were collected using custom Perl and Python scripts.
Data analysis	We used Python3, Jupyter notebooks and PyTorch to create and analyze models. The full list of libraries with version numbers is included in the code repository.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The imaging data are not publicly available because they contain confidential information that may compromise patient privacy as well as the ethical or regulatory policies of our institution. These data will be made available to individuals who contact the author Joseph Lo (joseph.lo@duke.edu) with reasonable requests for non-commercial research. Data usage agreements may be required.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No sample size calculation was performed. We used all available annotated mass cases, 1136 images across 484 patients. The test-validation-train split is standard for a small dataset in machine learning.
Data exclusions	Images were excluded if a biopsy clip was visible in the image because a biopsy clip could provide lesion localization information.
Replication	We trained both our models and baseline models using different random seeds, and found the same findings.
Randomization	We randomly assigned images to the test, train or validation set, such that the class balance was the same in each of the test, train and validation sets. We ensured that within each class there was no patient overlap between the test set and other sets.
Blinding	Blinding wasn't relevant to our study because it was a retrospective study.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	We did not differentiate patients by age, gender, genotypic information or previous treatment. The average patient age was 56.1 +/− 12.3 years.
Recruitment	Patients who received biopsies at Duke Hospital and Duke Hospital systems between 2008 and 2018 were selected. The relevant self-selection bias here is (1) which patients choose to receive regular mammographic scans, and (2) which patients do so at Duke and Duke Hospital systems.
Ethics oversight	Duke Health IRB

Note that full information on the approval of the study protocol must also be provided in the manuscript.