Explainable Artificial Intelligence for Breast Tumour Classification: Helpful or Harmful

Explainable Artificial Intelligence for Breast Tumour Classification: Helpful or Harmful

Amy Rafferty[1][0000−0002−5120−966X], Rudolf Nenutil[2], and Ajitha Rajan[1]

[1] University of Edinburgh, 10 Crichton St, Edinburgh EH8 9AB
[2] Masaryk Memorial Cancer Institute, Žlutý kopec 543/7, 602 00 Brno-střed-Staré Brno, Czechia

Abstract. Explainable Artificial Intelligence (XAI) is the field of AI dedicated to promoting trust in machine learning models by helping us to understand how they make their decisions. For example, image explanations show us which pixels or segments were deemed most important by a model for a particular classification decision. This research focuses on image explanations generated by LIME, RISE and SHAP for a model which classifies breast mammograms as either benign or malignant. We assess these XAI techniques based on (1) the extent to which they agree with each other, as decided by One-Way ANOVA, Kendall's Tau and RBO statistical tests, and (2) their agreement with the diagnostically important areas as identified by a radiologist on a small subset of mammograms. The main contribution of this research is the discovery that the 3 techniques consistently disagree both with each other and with the medical truth. We argue that using these off-shelf techniques in a medical context is not a feasible approach, and discuss possible causes of this problem, as well as some potential solutions.

Keywords: Machine Learning · Breast Tumour Classification · Explainable AI · LIME · RISE · SHAP

1 Introduction

Recent developments in deep learning (DL) have sparked an interest in more high-stakes applications such as medical diagnostics. Given a medical scan, a clinician may want to differentiate between healthy and unhealthy tissue, or between pathologies. However, the black-box nature of DL models means their conclusions tend not to be trusted by clinicians who cannot determine how the model came to its decision. Interpretable explanations are therefore crucial. Many medical experts have already expressed their concerns over rising black-box DL approaches [8].

Explainable AI (XAI) techniques exist to bridge this gap by intuitively highlighting the most important features of an input. This gives the model practitioner more information about how to improve the model's correctness, and gives the end-user, potentially a non-expert, an idea of how the model came to its conclusion. Knowing that a model's conclusion is correct is essential in medical diagnostics as their outcomes could impact lives.

Using XAI techniques for medical diagnostics comes with its own set of problems. Medical datasets are problematic due to differing labelling standards – some images have complete clinical annotation, while others simply state whether a tumour is present. Many techniques run into problems for images with small regions of interest (ROIs), due to their usage of image segmentation. This is the case for breast mammograms as cancerous regions can be extremely small. [23] discusses the serious implications of bad explanations in high stakes contexts. Saliency maps, which are commonly used to visualise image explanations, can be virtually identical for different classes on the same image [2]. Unreliable and misleading explanations can have serious negative implications.

We present a case study which focuses on the quality of explanations from 3 widely used XAI techniques, applied to a publicly available CNN-based classification model used

to identify malignant and benign breast tumours (originally designed for brain tumour detection [15]), and a public anonymised dataset of benign and malignant breast mammograms [12]. We assess the XAI techniques based on (1) the extent to which they agree with each other for the whole dataset, and (2) evaluation by two independent radiologists on the correctness of the important regions identified by each of the XAI techniques for 10 mammograms. The XAI techniques used in our study, LIME [20], SHAP [14] and RiSE [18], are discussed in the next Section.

2 Related Work

Many existing XAI techniques are applicable to the medical context. [31] presents an exhaustive list of techniques used for medical image analysis - we limit our consideration here to LIME, SHAP and RISE due to their popularity and ease of use [5]. We plan to consider other XAI techniques in the future.

LIME - Local Interpretable Model-Agnostic Explanations. LIME [20] is an XAI technique which can be applied to any model without needing any information about its structure. LIME provides a local explanation by replacing a complex neural network (NN) locally with something simpler, for example a Ridge regression model. LIME creates many perturbations of the original image by masking out random segments, and then weights these perturbations by their 'closeness' to the original image to ensure that drastic perturbations have little impact. It then uses the simpler model to learn the mapping between the perturbations and any change in output label. This process allows LIME to determine which segments are most important to the classification decision – these segments are then shown in the visual explanation output.

RISE - Randomized Input Sampling for Explanations of Black Box Models. RISE [18] works by first generating many random masks of an image, multiplying them elementwise with the image, and then feeding them directly into the original model for label prediction. Saliency maps are generated from a linear combination of the masks where weights come from the output probabilities predicted by the model. These saliency maps highlight the most important pixels of the image regarding its classification. This makes RISE extremely interpretable. RISE is also model agnostic. We note that RISE is very similar to LIME, however it measures saliency based on individual pixels, rather than superpixels, and therefore may perform better on images with small ROIs (eg. mammograms).

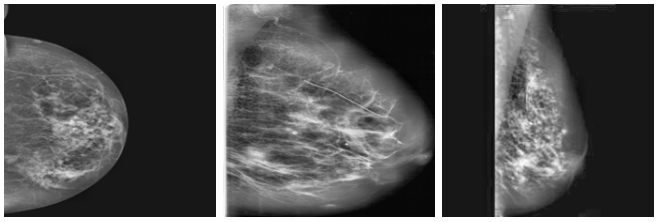
SHAP - Shapley Additive Explanations. SHAP [14] is another model-agnostic approach which uses Shapley values, a concept from game theory, to find the contribution of each feature to the model's output. The image is segmented to reduce the number of value computations. Starting from one random segment, we add one segment at a time until the correct model classification is possible. This is repeated many times with random orderings to get the importance of each segment, represented as Shapley values. Large positive SHAP values indicate that the segment is very important to the classification decision. SHAP is also a highly interpretable technique. We note that SHAP values are derived from game theory's Shapley values – they are not the same, and the mathematical differences are discussed in detail in [14].

2.1 XAI in Medicine

These methods, as well as other techniques [33] [27] [34] [26] [24] [29] [30], have had huge success, particularly in the image classification and Natural Language Processing fields,

however they are only beginning to be evaluated in any medical context [31]. An important issue to note is that when using larger medical images such as MRI scans, there is a need to split the images into tiles due to their extremely high resolutions. XAI techniques then need to be run on each tile, and the results need to be brought back together. Since we are working with mammograms, this is not an issue for this research, but is something we plan to explore in future work.

[25] highlights some of the challenges faced by medical professionals regarding XAI – not all visualisations are interpretable, there is no current definition for sufficient explainability in the field, and XAI techniques are not satisfactorily robust [1]. They describe the issue of the knowledge gap between AI and medical professionals, and the effects this has on techniques. Currently the focus of medical XAI seems to be on diagnosing rare diseases and monitoring health trends [25]. Some contributions to XAI for tumour classification exist, for example [6] which focuses on sequencing gene data, and [21] which also focuses on mammograms, though with gradient-based XAI techniques. [11] argues that explanations generated by LIME and SHAP cause no improvement on human decision making abilities – when shown an image both with and without an explanation, there was no statistical difference in the time it took for people to classify the image by eye, or in the number of mistakes made. This is concerning as the goal of diagnostic XAI is to make the lives of medical professionals easier, and remove the need for tedious by-eye classification [19] [10].



(a) INbreast (b) DDSM (c) MIAS

Fig. 1: Example images from each of the three dataset sources.

3 Model Setup

3.1 Data Pre-Processing

For this research we take breast mammograms with cancerous masses from a public dataset [12], which takes images from 3 official datasets – INbreast [16], DDSM [3] and MIAS [28]. When generating this dataset, the creators [13] extracted a small number of images with masses from each source, and performed data augmentation in the form of image rotation to generate a larger dataset. They also re-sized images to 227x227 pixels.

The public dataset [12] we are using is large. The original paper introducing this dataset [13] details their data augmentation techniques, which includes rotating and flipping each image to generate 14 variations of itself. This is not useful for this research – we are not trying to train a model that can cope with rotated breast scans, as the original scans and therefore any unseen real-world scans are all of the same orientation. We only use images of the original orientation. We also only take images from INbreast and DDSM, as the only MIAS scans present in the dataset were benign, though we plan to include MIAS for evaluation purposes (eg. Out-Of-Distribution detection (OOD) [17]) in future work to improve model confidence. The visual difference in original scans between the 3 sources is

shown in Figure 1. After selecting the images of the same orientation from the INbreast and DDSM sections of the dataset, we have a dataset of 2236 images - 1193 benign and 1043 malignant.

Image Cropping. For maximal model performance, we crop out as much of the black background as possible, making the breast the focus of each image. This was performed using basic Python opencv code. Images are then resized to the original 227x227 pixel format for consistency.

Dataset Split. Our dataset of 2236 images is split into a (Training / Validation / Testing) ratio of (2124 / 56 / 56). The Validation set will be used for all intermediate experiments – deciding how many epochs to train the model for, and tuning parameters for LIME. A small test set was chosen to ensure sufficient model training due to the small dataset size.

3.2 Model Architecture

We use an existing public CNN [15] which was originally used for binary classification of brain scans regarding the presence of a tumour. We use this model as it was specifically designed for the domain of tumours in medical scans, and was therefore reliable in the sense that it was likely to perform well on data like ours - noisy black and white scans containing cancerous legions. In the original study the model achieved 88.7% accuracy on the test set. The model takes an image and outputs a decimal value between 0 and 1, where 0 is benign, and 1 is malignant. We have taken 0.5 to be the threshold value for these classifications. The CNN contains 8 layers, using ReLU activation.

To avoid overfitting, we train four models differing only in numbers of epochs, and evaluate their performances on the Validation set. The performances of these models are described in Table 1 (Appendix A.1) in the form of their Accuracy and F1 Score. These statistics are based on a Validation set of 56 images. We proceed with the 75 epoch model as it has the highest performance scores. On the Test set, the 75 epoch model has an Accuracy of 0.9643 and an F1 Score of 0.9642. For training, we use Keras with the adam optimizer and binary cross-entropy loss function. We use a Intel(R) Core(TM) i5-7300HQ CPU @ 2.50GHz processor laptop for our experiments.

Although the accuracy of the model on the test set is high (96.43%), it is not clear whether the model infers classifications using the correct image features. In medical diagnostics, it is imperative that a clinician is able to interpret and understand the reasons for the classification label. Explanations from XAI techniques are meant to address this need.

4 Explanations

We generated individual explanations using each of the 3 XAI techniques, for a test set of 56 images. For illustration, we show explanations for the same 6 benign and 6 malignant examples with each XAI technique in Appendix A.8 (LIME Figure 6, RISE Figure 7, SHAP Figure 8). Code associated with generating explanations can be found at https://anonymous.4open.science/r/EvaluatingXAI-11DF/.

4.1 LIME

Our Python code for generating LIME explanations follows the steps described by [20]. For image segmentation, we used Python's scikit-image quickshift algorithm with empirically chosen parameters. When generating explanations, we highlight the boundaries of the L most important features for visibility. L was empirically chosen.

Choosing Segmentation Parameters For segmentation we use the scikit-image quickshift algorithm, which has 3 tuneable parameters – kernel size, max-dist, and ratio. These parameters and their effects are detailed in their documentation [7]. We use a small kernel size of 2, a default max-dist value of 10, and a small ratio value of 0.1. This was because we wanted many small segments with little emphasis on colour boundaries, to ensure that we consider small regions of interest (ROIs), and do not quantify the pixels at the boundary of the breast as incorrectly important.

Choosing L We define L as the number of most important features used in LIME explanations. The L value will determine the features shown in our LIME explanations, and also how many pixels are compared in the later One-Way ANOVA analysis. We will be comparing lists of most important pixels as decided by each XAI technique - the lengths of these lists will be equal for the 3 techniques, and will be the number of pixels within the L most important LIME features for a given image. We will then calculate the % pixel agreement between each pair of methods, defined as the proportion of pixels the lists have in common. To determine our L value, we calculate the average % pixel agreements between methods using L values of 3, 4, 5, 6 and 7. Averages are taken over the first 30 images in the Validation set for the sake of time. The results are shown in Figure 4 (Appendix A.2). As L increases, average pixel agreement increases between each pairwise technique comparison. It is infeasible to keep increasing L as we are trying to compare only the most important pixels. We have chosen L to be 6 as the first decrease in average agreement between all three techniques occurs at L = 7. Also, in the case of the pairwise comparisons LIME-SHAP and LIME-RISE, the jump in agreement from 6 to 7 is much smaller than from 5 to 6.

Observations. Figure 6 (Appendix A.8) shows 12 examples of LIME explanations - 6 for benign scans and 6 for malignant. For both classes, some explanations highlight undesirable features such as the image background. This is likely due to the variance in breast shape throughout the dataset, which can clearly be seen in these examples. This effect could be reduced by using larger datasets in the future. Looking at these explanations without ground truth tells us little about whether they are highlighting genuine cancerous regions. To evaluate LIME's performance, we compare its outputs to those of RISE and SHAP, and to a radiologist's evaluation.

4.2 RISE

Our Python code for RISE follows the steps described by [18]. Figure 7 (Appendix A.8) shows 12 examples of RISE explanations - 6 for benign scans and 6 for malignant. In these heatmaps, the most important pixels are shown as red, and the least important are shown as blue. We note that images have different importance value scales.

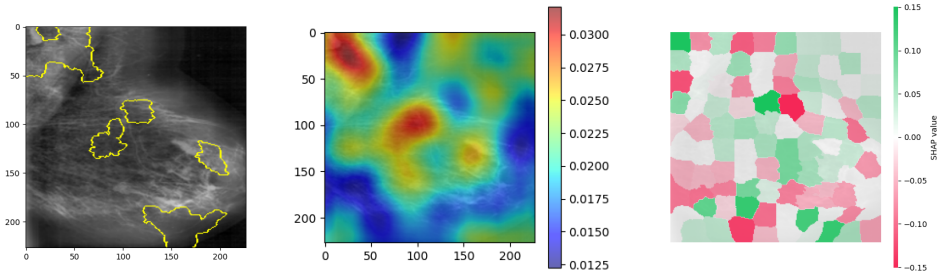
Observations. RISE generally assigns background pixels a medium relative importance. We expect that this is again due to irregular breast shapes. LIME and RISE seem to generate poor results for the same images - we define poor results as explanations which highlight background regions as important. Figure 7 (j) shows a case where RISE performs poorly for a malignant scan. LIME also performs poorly on this image, shown in Figure 6 (j). This image has an irregular shape, which supports our thoughts. Figure 7 (c) and Figure 6 (c) show the same issue for a benign scan.

4.3 SHAP

Our SHAP explanation code follows the steps described by [14]. Default values were used for image segmentation, and SHAP's Kernel Explainer was used. Figure 8 (Appendix A.8)

shows 12 examples of SHAP explanations - 6 for benign scans and 6 for malignant. Segments that contribute the most to the classification of the image are shown as green. The least important segments are shown as red. We note that SHAP value scales are not consistent across images.

Observations. Figure 8 (j) shows that SHAP performs poorly for the 4th malignant scan, much like LIME and RISE - heavily influential segments exist at the top-left corner of the image, which are background pixels. In most cases, the superpixels outside the boundary of the breast seem to have low SHAP values. SHAP seems to generally disregard background pixels with more success than RISE.



(a) LIME (b) RISE (c) SHAP

Fig. 2: Explanations by LIME, RISE and SHAP for a benign mammogram.

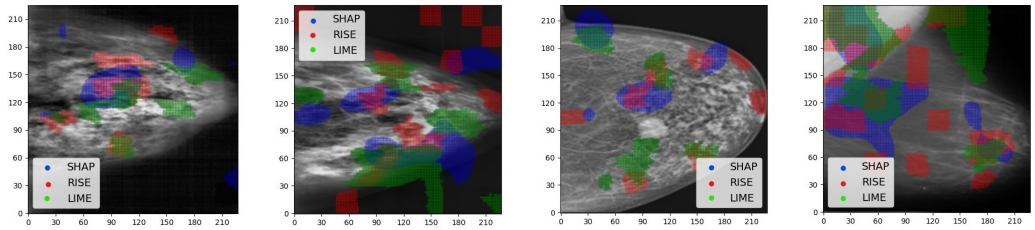
5 Evaluating Explanations

Looking at the 3 explanations side-by-side for an image, as in Figure 2, we can start to infer some agreement. However, due to the different explanation formats between techniques, the amount of agreement is unclear. In addition to visualisations, we use statistical analysis to compare the importance rankings of pixels between XAI techniques.

Visualising Agreement We use the 6 most important features in our LIME explanations, and denote n to be the number of pixels within these features. We visualise overlap between the n most important pixels given by each XAI technique, as in Figure 3. Generally, there are always areas which all 3 techniques identify as highly important. However there are more regions where they disagree. Sub-figures (b) and (d) from Figure 3 show cases where explanations have performed poorly - defined as identifying background pixels as most important. This is likely due to irregular breast shapes within the dataset. Figures 3 (a) and (c) show cases where explanations have multiple clear points of agreement.

5.1 One-Way ANOVA

One-Way ANOVA [22] compares the means of two or more groups for a dependent variable. Our input groups are 3 lists of % pixel agreements between methods for each test set image, labelled LIME-RISE, LIME-SHAP, and RISE-SHAP. To generate these lists, we identify



(a) Benign 1 (b) Benign 2 (c) Malignant 1 (d) Malignant 2

Fig. 3: Overlap between the 3 XAI techniques regarding the n most important pixels. SHAP is blue, RISE is red, LIME is green.

the n most important pixels according to each technique, where n is the number of pixels within the top 6 LIME features for a given image. This is because LIME outputs binary values for each pixel (presence in the L most important features) while RISE and SHAP assign decimal importance values. We then calculate the % pixel agreement across each pair of pixel lists for each image. We define % pixel agreement as the proportion of pixels the lists have in common. Results are in Table 2 in Appendix A.3.

The only statistically significant test is Test 2, shown in Table 2. This tells us that of all pairwise comparisons, there is only statistically significant difference in average pixel agreement between the comparisons of LIME-RISE and RISE-SHAP. Analysing the statistical composition of the pixel agreement lists supports this conclusion. Figure 5 (Appendix A.4) visualises these results. The largest difference in mean (green triangles) is between LIME-RISE and RISE-SHAP. Figure 5 and Table 3 (in Appendix A.4) show that the average pixel agreement between techniques is startlingly low – 20 - 30% on pairwise comparisons, and under 10% when comparing all three. However, these values still represent significant numbers of pixels, as our images are large and have small ROIs.

5.2 Kendall's Tau

Kendall's Tau [9] is a measure of the degree of correlation between two ranked lists. The purpose of Kendall's Tau is to discover whether two ordered lists are independent. We perform this test using the built-in Python scipy method, and set the inputs to be the ordered lists of pixels and their importance values for each of the 3 XAI techniques, in the form "(x, y): value".

We apply Kendall's Tau to each test set image 3 times – on the full pixel list, on the top n most important pixels, and on the top 1000. We want to discover any statistically significant correlation regarding the most important pixels to the classification between techniques - if there is, and the Tau values are positive, this implies agreement. Results are shown in Table 5 (in Appendix A.6). We use an alpha value of 0.05. From Table 5 we can conclude that the only instances of statistically significant correlation come from the LIME-SHAP comparison - both on the full length pixel list and the top n pixels. Positive Tau values imply a positive correlation. The LIME-RISE comparison yields results closer to the 0.05 threshold while RISE-SHAP yields the worst results. In Figure 5, we saw that for the top n pixel lists, LIME and RISE have the highest mean pixel agreements. This implies that while LIME and RISE have higher pixel agreement regarding the presence of the same pixels in the top n pixel lists, LIME and SHAP agree the most regarding pixel order.

We also evaluated our explanations using the RBO statistical test [32] to compare pixel rankings. The results of this test are shown and discussed in Appendix A.5.

5.3 Radiologist Evaluation

To assess our explanations with respect to the medical truth as understood by a clinician, we consulted 2 independent radiologists and provided them with a subset of 10 images - 5 with benign and 5 with malignant classification, each of them associated with explanations from the 3 different techniques. We were unable to gather an expert evaluation for the entire test set due to limited availability of the radiologists, though we intend to expand this form of XAI technique evaluation in future work.

The results gathered from this evaluation with 10 images for the first and second radiologists are shown in Tables 6 and 7 (in Appendix A.7). In these tables,'B' in the column heading refers to a benign image and 'M' refers to a malignant image. We requested the radiologists to score each explanation between 0 and 3 to represent its agreement to radiologist identified image regions. The definition of the scores provided to the radiologists are as follows:

0 = Explanation completely differs from expert opinion
1 = Explanation has some similarities, but mostly differs from expert opinion
2 = Explanation mostly agrees with expert opinion, though some areas differ
3 = Explanation and expert opinion completely agree

It is worth noting that no explanation earned a label of 3 from either radiologist – each explanation either identified erroneous regions or missed important sections. LIME appears to perform the worst within this subset of 10 images, while RISE performs the best. This is likely because RISE is the only method which uses pixels rather than superpixels and is therefore more fine grained when examining image regions and less likely to miss small regions of interest. There does not seem to be any difference in explanation quality between benign and malignant images for any technique.

The radiologists noted the following limitations with the explanation techniques:
- None of the explanations could identify the entire tumour region. Explanation methods only highlight fragmented relevant regions and this is along with many irrelevant regions.
- Explanations for both malignant and benign tumours are distributed all over the image and fail to take into account clinical features, like shapes of masses, margins, the density of tissues, and structural distortion.

Our radiologist evaluation using 10 mammograms may not be representative of a real world dataset. However, the issues highlighted by these comments are consistent problems - this will be discussed in Section 6.

5.4 Threats to Validity

This research uses a small dataset of breast mammograms which may not be representative of the population. We limit our classification task to benign or malignant - in reality there are many types of lesion for both classes, which would appear differently in mammograms. In future work, using a non-binary classifier alongside a more thorough radiologist evaluation may allow us to better analyse the failures of our techniques. We have assumed that the low cohesion between XAI techniques paired with the high model test accuracy indicates that failures are due to the XAI techniques, and not the model itself. In future work we will utilise multiple models and explore alternate XAI evaluation techniques [4] in order to back up this claim. All empirical analysis regarding LIME parameter tuning and the choice of L was based solely on the patterns within our data. They may not hold up when compared to a larger dataset. Our XAI techniques by definition utilise randomization when generating masks, therefore re-running our code will generate slightly different results to the ones displayed here. This variation is not hugely impactful as we generally discuss average

values in our statistical tests. Our code for LIME, RISE and SHAP is not the only way of implementing these techniques - there are many public examples which implement the steps described in the literature in slightly different ways. Because of this, another researcher's code may yield different results to the ones shown here.

6 Observations and Discussion

Each technique performs poorly on the same images. Our explanations highlight the quality variation within the test set. Each XAI technique performed poorly (highlighted background pixels as most important) on the same images, usually mammograms with irregular breast shapes. This is likely due to our small dataset and the effect of blurring and image re-sizing. It's interesting to note that these problems don't seem to impede the model accuracy, only the quality of explanations.

Percentage pixel agreement between XAI techniques is extremely low. LIME and RISE appear to have the most pixel agreement according to One-Way ANOVA. However, these values are not high, with an average agreement of 28%. Combining Kendall's Tau with One-Way ANOVA, we find that while LIME and RISE consistently highlight the highest proportion of the same important pixels, LIME and SHAP have the most similar pixel orderings. This is supported by RBO.

Radiologist evaluation revealed explanations from all three techniques were unhelpful. The radiologists found that RISE performed marginally better than the other two techniques. Explanations from all three techniques, however, do not consider clinical features within mammograms that are used to diagnose benign or malignant tumours, such as shape of mass, boundary, and density. The explanations do not highlight the entire tumour as important, but instead sparsely pick parts of the tumour along with many irrelevant regions. The XAI techniques we have used have low levels of agreement with each other, as well as low levels of agreement with the medical truth.

6.1 Discussion

The goal of this research was to determine whether taking off-shelf XAI techniques and applying them to breast tumour classification was a feasible approach that would hold up in the real world. Bringing together our observations tells us that this is not the case.

Though LIME and SHAP have the highest agreement in pixel orderings, these agreement levels are still very low. Explanations from these techniques highlight some common areas, though have significant disagreements and are therefore unreliable for use in diagnostics. The most likely reason that LIME and SHAP have the highest pixel ordering agreement is that these methods both utilise superpixels, while RISE does not. Discussing similarities in pixel orderings is problematic in this context, due to the differing ways in which each of the 3 XAI techniques assign importance values to pixels. We note that these differences come from both the underlying properties of each technique, and from our code architecture. LIME's binary scoring method is likely the reason behind the slightly higher % pixel agreement statistics for pairwise comparisons involving LIME.

Each XAI technique works differently, and resulting explanations depend on many different factors – segmentation, mask randomization, and tuneable parameters. While this is an expected reason for some result variation, a higher level of cohesion in explanations was to be expected. We identified that each technique incorrectly highlighted background regions as being most important on images with irregular breast shapes. While this may have been caused by the small size of the dataset, and image quality after pre-processing, we would

have expected the model's accuracy to also decline to reflect this, and it did not. We also note that the techniques showed no difference in explanation quality for images from the benign or malignant classes.

Regarding the medical truth according to a radiologist, RISE seems to produce the most medically correct explanations, while the results of LIME and SHAP are often entirely incorrect. This is likely because RISE involves no image segmentation. No explanations were labelled as perfect - areas are always missed or incorrectly highlighted. We therefore conclude that explanations generated by LIME, RISE and SHAP are in disagreement with respect to both each other, and to the medical truth, and so do not perform reliably in this context. The results of these explanation techniques do not match or consider what a radiologist would want in a real-world context. Instead of pixels or superpixels, techniques should identify clinically defined regions. This is a gap that needs to be bridged - we highlight the need for specific, carefully defined techniques for explaining tumour images that take clinical features into account.

References

1. Alvarez-Melis, D., Jaakkola, T.S.: On the robustness of interpretability methods (2018), https://arxiv.org/abs/1806.08049
2. Arun, N., Gaw, N., Singh, P., Chang, K., Aggarwal, M., Chen, B., Hoebel, K., Gupta, S., Patel, J., Gidwani, M., Adebayo, J., Li, M.D., Kalpathy-Cramer, J.: Assessing the (un)trustworthiness of saliency maps for localizing abnormalities in medical imaging (2020), https://arxiv.org/abs/2008.02766
3. Heath, M.D., Bowyer, K., Kopans, D.B., Moore, R.H.: The digital database for screening mammography (2007)
4. Hedström, A., Weber, L., Bareeva, D., Motzkus, F., Samek, W., Lapuschkin, S., Höhne, M.M.C.: Quantus: An explainable ai toolkit for responsible evaluation of neural network explanations (2022), https://arxiv.org/abs/2202.06861
5. Holzinger, A., Saranti, A., Molnar, C., Biecek, P., Samek, W.: Explainable AI Methods - A Brief Overview, pp. 13–38. Springer International Publishing (2022)
6. Huang, L.: An integrated method for cancer classification and rule extraction from microarray data. J Biomed Sci 2009 **16 (1): 25** (2009)
7. scikit image.org: Scikit-image documentation. https://scikit-image.org/docs/stable/api/skimage.segmentation.html
8. Jia, X., Ren, L., Cai, L.: Clinical implementation of AI techniques will require interpretable AI models. In: Med. Phys. 47. pp. 1 – 4 (2020)
9. Kendall, M.: A new measure of rank correlation. In: Biometrika 30. pp. 81 – 89 (1938)
10. King, B.: Artificial intelligence and radiology: what will the future hold? vol. 15(3 Part B), pp. 501–503 (2018)
11. Knapič, S., Malhi, A., Saluja, R., Främling, K.: Explainable artificial intelligence for human decision support system in the medical domain. Machine Learning and Knowledge Extraction (2021)
12. Lin, T., Huang, M.: Dataset of breast mammography images with masses. In: Mendeley Data, V5 (2020)
13. Lin, T., Huang, M.: Dataset of breast mammography images with masses. In: Data in Brief, Volume 31, 105928 (2020)
14. Lundberg, S., Lee, S.: A unified approach to interpreting model predictions. In: Proceedings of the 31st international conference on neural information processing systems. pp. 4768 – 4777 (2017)
15. MohamedAliHabib: Brain tumour detection, Github repository. In: https://github.com/MohamedAliHabib/Brain-Tumor-Detection. GitHub (2019)
16. Moreira, I., Amaral, I., Domingues, I., Cardoso, A.J.O., Cardoso, M.J., Cardoso, J.S.: Inbreast: toward a full-field digital mammographic database. Academic radiology **19 2**, 236–248 (2012)
17. Park, J., Jo, K., Gwak, D., Hong, J., Choo, J., Choi, E.: Evaluation of out-of-distribution detection performance of self-supervised learning in a controllable environment (2020). https://doi.org/10.48550/ARXIV.2011.13120, https://arxiv.org/abs/2011.13120

18. Petsiuk, V., Das, A., Saenko, K.: RISE: Randomized input sampling for explanation of black-box models. In: arXiv, 1806.07421 (2018)
19. Recht, M., Bryan, R.: Artificial intelligence: threat or boon to radiologists? vol. 14(11), pp. 1476–1480 (2017)
20. Ribeiro, M., Singh, S., Guestrin, C.: "Why should I trust you?": Explaining the predictions of any classifier. In: arXiv, 1602.04938v3 (2016)
21. Rodriguez-Sampaio, M., Rincón, M., Valladares-Rodriguez, S., Bachiller-Mayoral, M.: Explainable artificial intelligence to detect breast cancer: A qualitative case-based visual interpretability approach. In: Ferrández Vicente, J.M., Álvarez-Sánchez, J.R., de la Paz López, F., Adeli, H. (eds.) Artificial Intelligence in Neuroscience: Affective Analysis and Health Applications. pp. 557–566. Springer International Publishing (2022)
22. Ross, A., Willson, V.L.: One-Way Anova, pp. 21 – 24. SensePublishers, Rotterdam (2017)
23. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. In: Nat Mach Intell 1. pp. 206 – 215 (2019)
24. Selvaraju, R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: arXiv, 1610.02391 (2017)
25. Seyedeh, P., Zhaoyi, C., Pablo, R.: Explainable artificial intelligence models using real-world electronic health record data: a systematic scoping review. Journal of the American Medical Informatics Association **27**, 1173 – 1185 (2020)
26. Shrikumar, A., Greenside, P., Kundaje, A.: Learning important features through propagating activation differences (2017), https://arxiv.org/abs/1704.02685
27. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualizing image classification models and saliency maps. https://arxiv.org/abs/1312.6034 (2014)
28. Suckling, J., Parker, J., Dance, D.: Mammographic image analysis society (MIAS) database v1.21. In: https://www.repository.cam.ac.uk/handle/1810/250394 (2015)
29. Sun, Y., Chockler, H., Huang, X., Kroening, D.: Explaining image classifiers using statistical fault localization. In: European Conference on Computer Vision (ECCV) (2020)
30. Sun, Y., Chockler, H., Kroening, D.: Explanations for occluded images. In: International Conference on Computer Vision (ICCV). pp. 1234–1243. IEEE (2021)
31. van der Velden, B.H., Kuijf, H.J., Gilhuijs, K.G., Viergever, M.A.: Explainable artificial intelligence (xai) in deep learning-based medical image analysis. Medical Image Analysis p. 102470 (2022)
32. Webber, W., Moffat, A., Zobel, J.: A similarity measure for indefinite rankings. In: ACM Transactions on information systems. vol. 28, 4 (2010)
33. Zeiler, M., Fergus, R.: Visualizing and understanding convolutional networks. Fleet D, Pajdla T, Schiele B, Tuytelaars T, editors. Computer Vision–ECCV 2014; Zurich, Switzerland. pp. 818 – 833 (2014)
34. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: arXiv, 1512.04150 (2016)

A Appendix

A.1 Model Training Results

The results of the experiment used to choose the 75 epoch model when considering the impact of overfitting on our CNN, as discussed in Section 3.2 of this report.

Table 1: Validation Accuracy and F1 Score for CNNs trained with different numbers of epochs.

Epochs	Accuracy	F1 Score
25	0.8214	0.7917
50	0.8214	0.7917
75	**0.8750**	**0.8571**
100	0.8036	0.7660

A.2 Choosing L Parameter for LIME

The results of the experiment used to choose L, as discussed in Section 4.1 of this report.

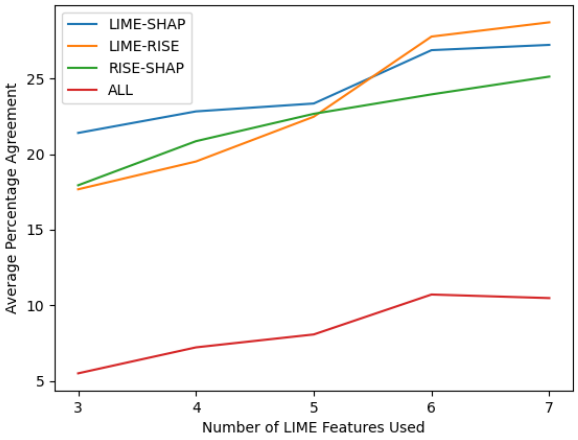


Fig. 4: Average % pixel agreement values between techniques taken over 30 images from the Validation set.

A.3 One-Way ANOVA Results

We present here the statistical hypotheses used for the One-Way ANOVA test, as well as the results gathered. This statistical test and its implications is discussed in Section 5.1 of this report. The results are shown in Table 2.

The hypotheses for One-Way ANOVA are as follows:
- H0: There is no statistically significant difference between the means of the groups.
- H1: There is a statistically significant difference between the means of the groups.

Table 2: Results of One-Way ANOVA tests as described in the text. **Bold** results are statistically significant (alpha value 0.05).

Test	Methods Compared	F-statistic	p-value
1	LIME-RISE, LIME-SHAP	3.7823	0.0544
2	LIME-RISE, RISE-SHAP	**9.1855**	**0.0031**
3	RISE-SHAP, LIME-SHAP	1.6193	0.2060

A.4 Pixel Agreement Statistics

Figure 5 presents a box plot representation of the % pixel agreement values between XAI techniques, taken over all images in our test set. These results are discussed in Section 5.1 of this report. Table 3 also represents these agreement values.

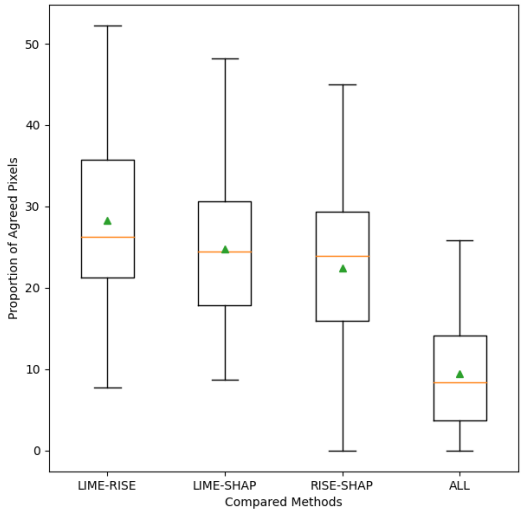


Fig. 5: % Pixel Agreement between techniques for n most important pixels. Medians are orange lines, means are green triangles.

A.5 Ranked Biased Overlap (RBO) Results

RBO [32] weights each rank position by considering the depth of the ranking being examined, minimising the effect of the least important pixels. Taking two ranked lists as inputs, RBO outputs a value between 0 and 1, where 0 indicates that the lists are disjoint, and 1 indicates that they are identical. The results of RBO depend on the tuneable parameter p [32]. Small p values place more weight on items at the top of an ordered list. While this is desirable, we must consider the difference in pixel importance value allocation methods between techniques. RISE applies a decimal score to each pixel. SHAP applies the same decimal score to each pixel within a given image segment. LIME uses binary values indicating whether the pixels are in the top 6 most important features. We use large p values to properly encompass similarities between larger groups of pixels with identical values.

Table 4 shows the average, minimum and maximum RBO values for each pairwise pixel list comparison. The average RBO values for each comparison tell us that the pixel lists are

Table 3: Statistical overview of percentage pixel agreements for all method comparisons.

Techniques	Mean	Std	Min	Max
LIME-RISE	28.27%	10.13%	7.74%	52.19%
LIME-SHAP	24.73%	8.75%	8.73%	48.16%
RISE-SHAP	22.45%	9.82%	0.00%	44.97%
ALL	9.48%	6.18%	0.00%	25.88%

almost disjoint. This is expected due to the differing pixel importance allocation methods as discussed. Instead we consider the maximum values – LIME and SHAP generate lists that are hugely identical for at least one instance in the test set, with maximum RBO values in the range 0.69 – 0.78. The other pairwise comparisons do not come close to these numbers. This observation supports Kendall's Tau – both tests have identified LIME and SHAP as the techniques with the highest agreement regarding pixel orderings.

Table 4: RBO results performed on full ordered pixel importance lists for each technique, with differing p values. Values shown to 3 decimal places, though we note here that these values are never exactly zero, just extremely small.

-	RISE-SHAP			LIME-SHAP			LIME-RISE		
p	0.9	0.95	0.99	0.9	0.95	0.99	0.9	0.95	0.99
Min	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Max	0.000	0.000	0.077	0.697	0.763	0.782	0.002	0.023	0.265
Avg	0.000	0.000	0.003	0.019	0.027	0.045	0.000	0.001	0.011

A.6 Kendall's Tau Results

We present here the statistical hypotheses used for the Kendall's Tau test, as well as the results gathered. This statistical test and its implications is discussed in Section 5.2 of this report. The results are shown in Table 5.

The following hypotheses are used:
- H0: There is no statistically significant correlation, the lists are independent.
- H1: There is a statistically significant correlation in pixel orderings between lists, they are not independent.

Table 5: Kendall's Tau comparison results. n is defined in the text. Values are averages taken over the test set, shown to 3 decimal places. **Bold** results are statistically significant.

Techniques	p-values			Tau		
	Full	n	1000	Full	n	1000
RISE-SHAP	0.123	0.125	0.249	0.003	0.002	0.001
LIME-SHAP	**0.000**	**0.048**	0.067	**0.154**	**0.106**	0.293
LIME-RISE	0.066	0.055	0.133	0.004	-0.006	0.014

A.7 Radiologist Opinions

Here we present the results as received from two independent radiologists, as well as definitions of the scoring system used to evaluate explanations,

We requested each explanation be scored between 0 and 3 to represent its agreement to radiologist identified image regions. The definition of the scores provided to the radiologists are as follows:

0 = Explanation completely differs from expert opinion
1 = Explanation has some similarities, but mostly differs from expert opinion
2 = Explanation mostly agrees with expert opinion, though some areas differ
3 = Explanation and expert opinion completely agree

Table 6: Radiologist evaluation regarding explanations generated on a subset of 10 images. B denotes benign, and M denotes malignant.

Image	B1	B2	B3	B4	B5	M1	M2	M3	M4	M5
LIME	0	1	0	1	1	0	0	1	0	0
RISE	0	1	1	1	1	2	1	1	1	2
SHAP	0	0	0	1	2	0	1	1	1	0

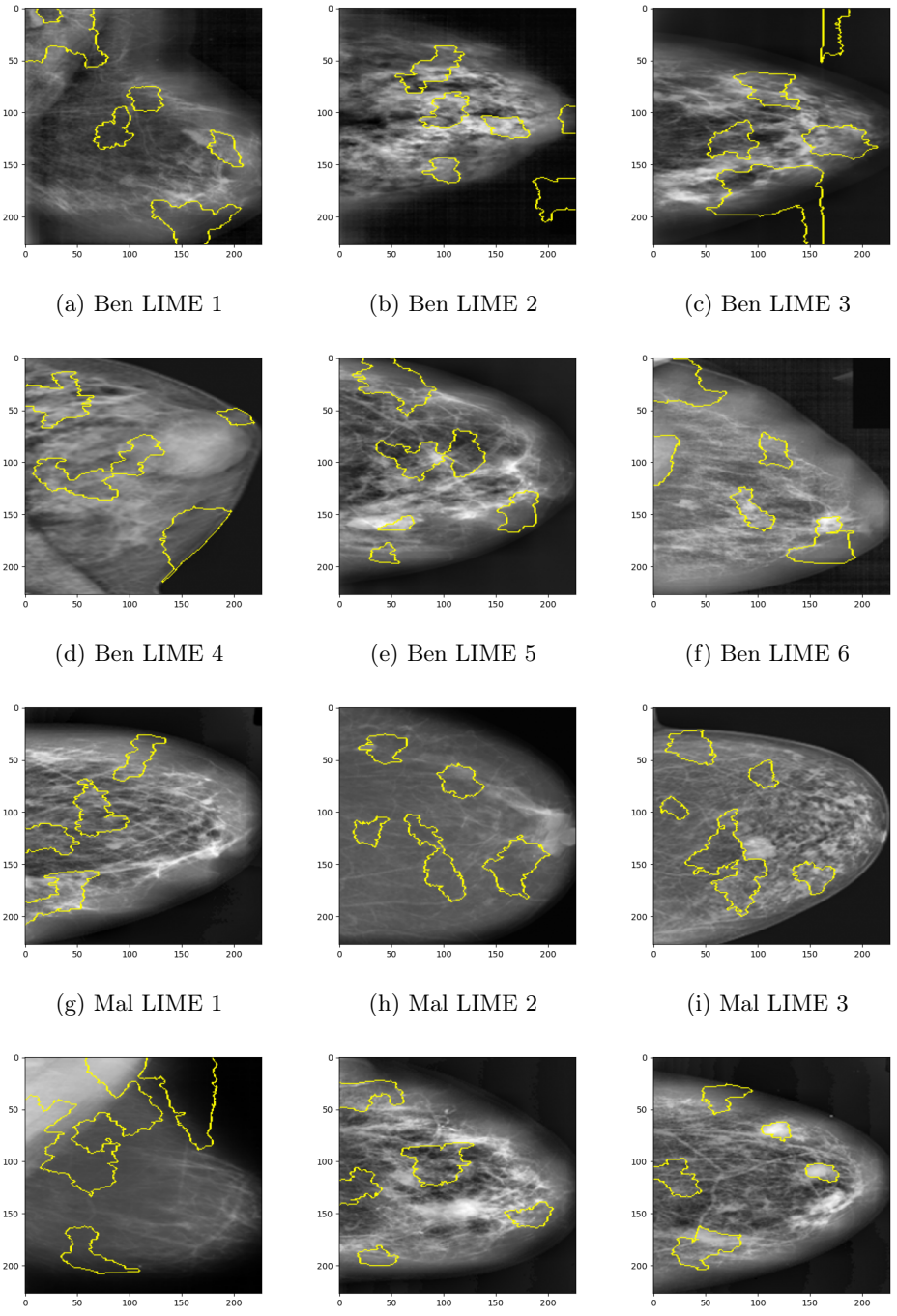
Table 7: Second radiologist evaluation regarding explanations generated on a subset of 10 images. B denotes benign, M denotes malignant.

Image	B1	B2	B3	B4	B5	M1	M2	M3	M4	M5
LIME	0	2	0	1	1	0	0	0	0	0
RISE	0	0	0	1	1	2	0	0	0	0
SHAP	0	2	0	1	0	1	0	1	1	1

We note that the opinions of the two radiologists above do not entirely agree with each other - this is due to the fact that identifying all cancerous regions by eye, especially on benign mammograms, is extremely difficult. The scans are also fairly noisy and in parts blurry by nature. The purpose of this form of evaluation was not to have radiologists perfectly highlight all cancerous regions - the goal was to simply analyse their responses to explanations generated by each XAI technique, in order to judge the usefulness of the techniques as diagnostic tools.

A.8 Explanation Examples

This section contains examples of image explanations as generated by LIME, RISE and SHAP, described in this report. Figure 6 shows LIME explanations, Figure 7 shows RISE explanations, and 8 shows SHAP explanations.



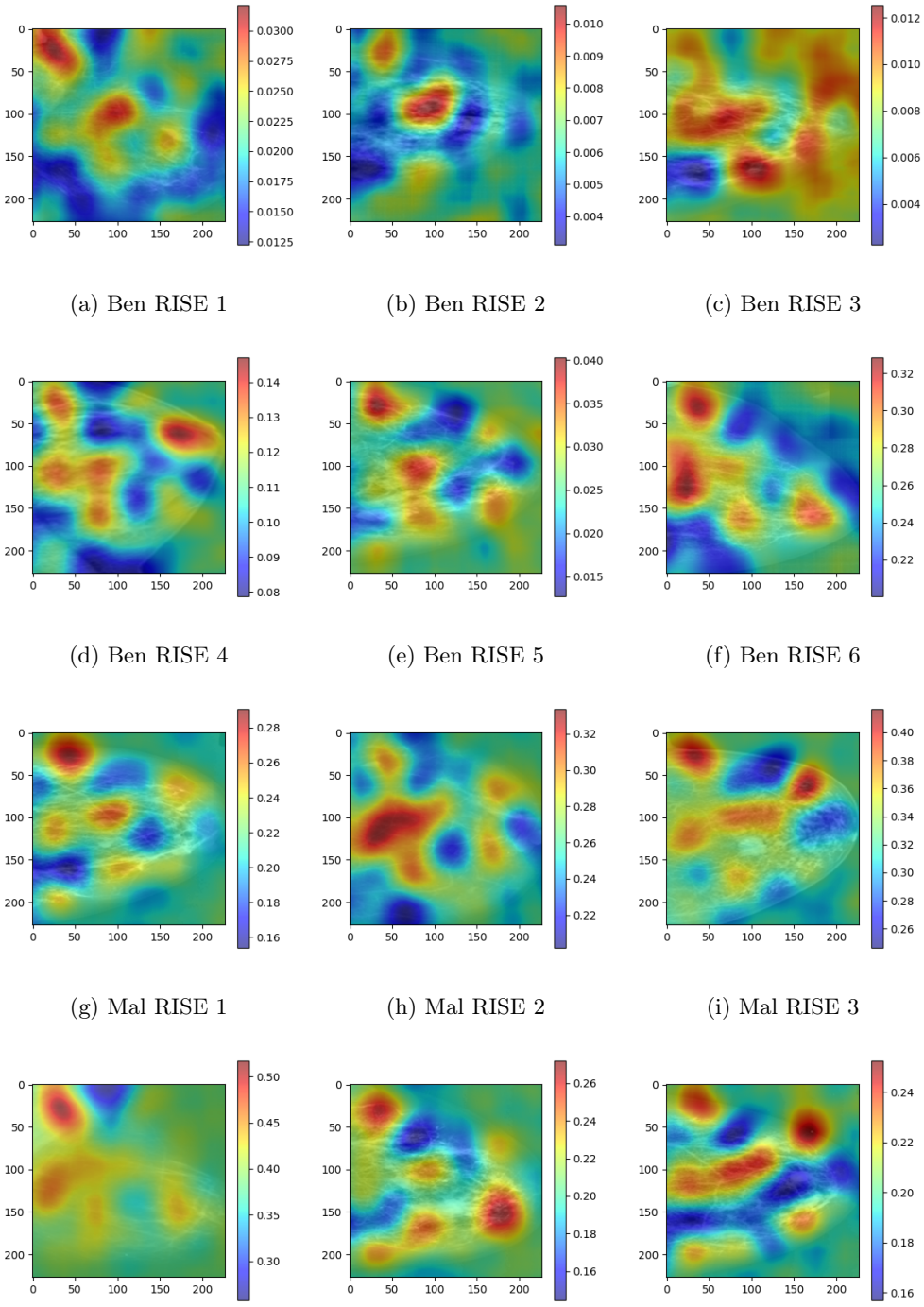
(a) Ben LIME 1 (b) Ben LIME 2 (c) Ben LIME 3

(d) Ben LIME 4 (e) Ben LIME 5 (f) Ben LIME 6

(g) Mal LIME 1 (h) Mal LIME 2 (i) Mal LIME 3

(j) Mal LIME 4 (k) Mal LIME 5 (l) Mal LIME 6

Fig. 6: Examples of LIME explanations generated for benign (Ben) and malignant (Mal) breast mammograms.



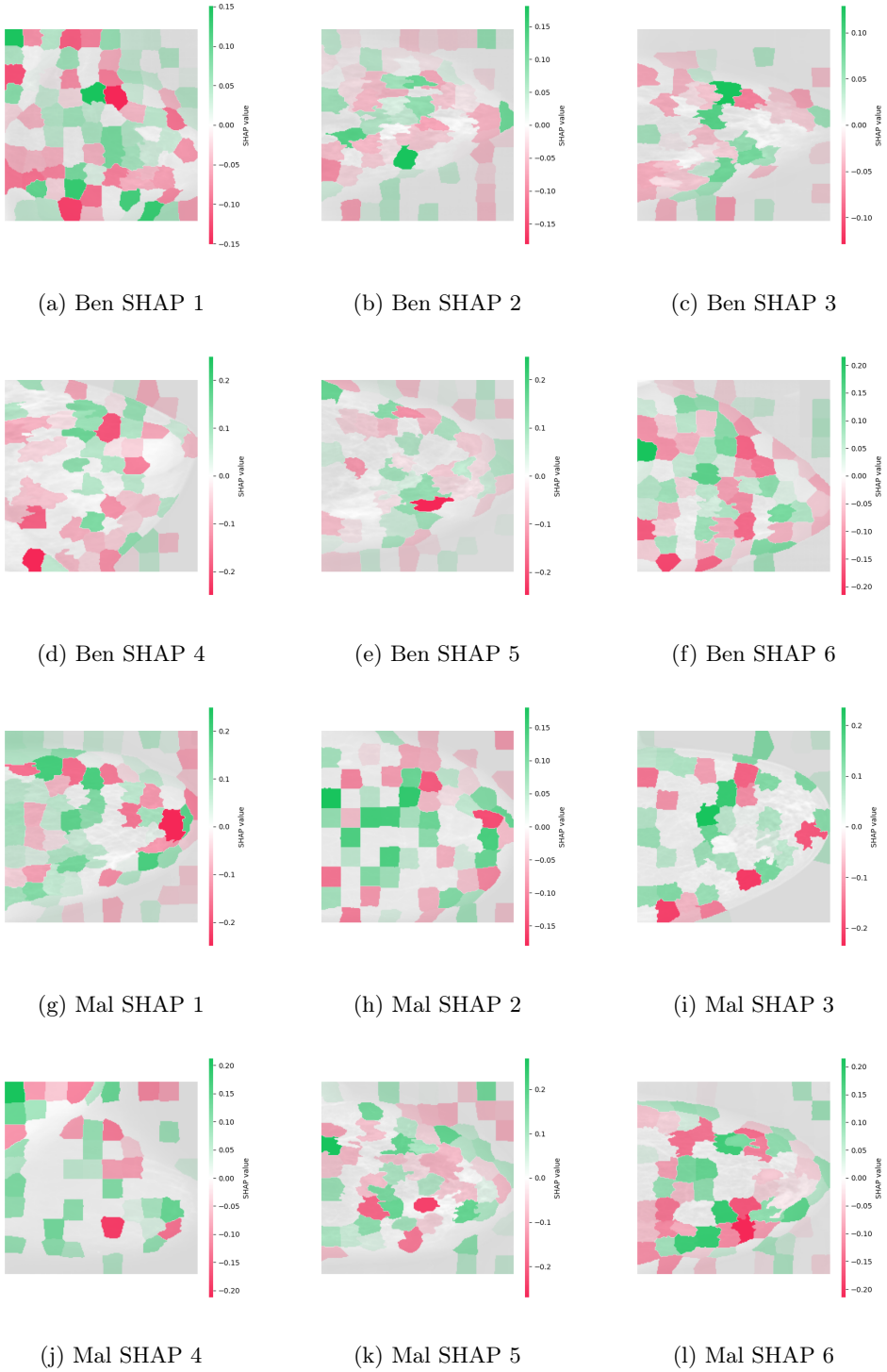
(a) Ben RISE 1 (b) Ben RISE 2 (c) Ben RISE 3

(d) Ben RISE 4 (e) Ben RISE 5 (f) Ben RISE 6

(g) Mal RISE 1 (h) Mal RISE 2 (i) Mal RISE 3

(j) Mal RISE 4 (k) Mal RISE 5 (l) Mal RISE 6

Fig. 7: Examples of RISE explanations generated for benign (Ben) and malignant (Mal) breast mammograms.



(a) Ben SHAP 1 (b) Ben SHAP 2 (c) Ben SHAP 3

(d) Ben SHAP 4 (e) Ben SHAP 5 (f) Ben SHAP 6

(g) Mal SHAP 1 (h) Mal SHAP 2 (i) Mal SHAP 3

(j) Mal SHAP 4 (k) Mal SHAP 5 (l) Mal SHAP 6

Fig. 8: Examples of SHAP explanations generated for benign (Ben) and malignant (Mal) breast mammograms.