# Evaluating Explainable Artificial Intelligence Techniques for Breast Tumour Classification

*Amy Rafferty*


4th Year Project Report
Artificial Intelligence and Computer Science
School of Informatics
University of Edinburgh

2022

# Abstract

Explainable Artificial Intelligence (XAI) is the field of AI dedicated to promoting trust in machine learning models by helping us understand how they make their decisions. Any ML model can take an input and produce an output – but how did the model reach this conclusion? How do we know it is correct? Can we trust the model in a high-stakes environment? These are the questions addressed by XAI. Explanations intuitively show us which aspects of the data were deemed the most important by the model for the classification decision - for example in the case of image data, an explanation may show pixel importance as a heatmap. An explanation for an animal identifier may highlight a cat's ears, or a dog's tail – in cases like this where we know what to expect, we can quickly judge whether a model is using the correct regions to inform its decisions. Knowing this allows us to trust the model, to know that it is correct – but what about for more complex images? What if we don't know which regions we expect the explainer to highlight?

This project focuses on explanations for a model which classifies breast mammograms into two categories – Benign and Malignant. Medical datasets are problematic by nature – different labelling standards, blurriness, and stringent data protection laws can impede our ability to gather a large, high-quality dataset, and train a ML model which produces a reliable conclusion. Medical diagnostics is a very high-stakes field – it is extremely important that a clinician, who may not have any understanding of machine learning, can trust any model they are working with, and understand exactly which parts of a medical scan contributed to the model's decision. We will apply 3 commonly used XAI techniques (LIME, RISE and SHAP) to a CNN trained on breast mammograms from a public and anonymised medical dataset. For each image, we will generate an explanation from each technique, and use visual and statistical analysis to determine whether these explanations agree with each other – whether they consistently highlight the same regions as being important to a diagnosis decision. We will also show a small subset of explanations to a radiologist to determine whether they agree with the medical truth. The statistical methods used are One-Way ANOVA, RBO and Kendall's Tau. If the 3 XAI techniques consistently highlight common regions, and agree with the opinions of a radiologist, we can infer that the explanations produced are correct. Thus we can agree that using these off-shelf XAI techniques on complex medical scans is a reasonable approach, and show that these XAI techniques would therefore be useful in real-world diagnostics.

The main contribution of this project, however, is the discovery that the 3 techniques consistently disagree, both with each other and with a radiologist. Taking the n most important pixels as decided by each technique, where n is defined during experiments, we find that on average each pair of techniques only has 20 – 30% of the same pixels present in these lists. We also compare each technique's pixel orderings using RBO and Kendall's Tau tests, and discover that lists of descending pixel importance are almost disjoint between methods. We therefore argue that using these off-shelf techniques in such a specific medical context is not a feasible approach, and discuss some possible causes for this problem, as well as some potential solutions.

# Research Ethics Approval

This project was planned in accordance with the Informatics Research Ethics policy. It did not involve any aspects that required approval from the Informatics Research Ethics committee.

All data is taken from a public and anonymised medical dataset.

# Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Amy Rafferty*)

# Acknowledgements

I would like to thank my supervisor Ajitha Rajan for organising this project, and for providing valuable feedback throughout each stage. I would also like to thank her for the opportunities she gave me to present my findings to various professionals, both in the medical and XAI fields, as this has been extremely useful for me to both improve my presentation skills and gain valuable real-world insight into the implications of my project.

I would also like to thank Ted Hupp and Rudolf Nenutil for their contributions as medical experts, as their comments about the correctness of my explanations regarding the medical truth really helped me bring my conclusions together.

Finally, I'd also like to thank my friends and my flatmate for keeping me sane, and my boxing coach Craig for putting up with me and stopping me from getting too stressed.

# Table of Contents

# Chapter 1

# Introduction

Artificial Intelligence is all around us. From self-driving cars, to evaluating loan suitability and medical diagnoses, AI systems are constantly making difficult and complex decisions that affect our day-to-day lives. Behind the scenes, Neural Networks (NNs) are the powerhouses by which these systems actually come to their conclusions. These NNs are complex functions involving different weight coefficients being applied to different parts (eg. pixels) of an input. It is very difficult to retrace the steps taken by these networks to form their outputs, and so we call them "black-box" models, owing to the fact that we cannot necessarily see their inner workings. However, the black-box nature of these systems leads to doubt and confusion as to how, exactly, a conclusion is reached, and whether it is correct or trustable within a high-stakes context. A self-driving car can decide that the road ahead is clear of pedestrians, and drive on, but what if it's wrong? Why should we trust the outcomes of these models, if we don't understand how these conclusions are being made? This is where the topic of model explanations becomes essential.

Explanations can be split into two types, differing in the group of people they intend to help, but having the same goals in mind. The first is to give the model practitioner more information about how to correctly utilise the model parameters for their own needs, improving the model's correctness and the likelihood of it being used in the real world. The second, and most common type, is to give the end-user, potentially a non-expert, an idea of how the model came to its conclusion. Knowing how a conclusion came to be helps us to first judge whether it is correct, and secondly trust the model that created it. This is very important in high-stakes fields such as medicine, as the outcomes of these black-box models could be life changing, and a mistake could therefore impact lives.

Figure 1.1 is taken from a review by Ras G. et al. (2020) [5] which demonstrates some intuitive examples of model explanations for different data types. These explanations highlight what the NN has decided to be the most important features when generating its prediction. For image data, these features are pixels. For tabular data, they are constraints, and for text data, they are words.

Explanations are particularly useful regarding image classification, as the end-user can see exactly which pixels have contributed most to a particular classification. This could

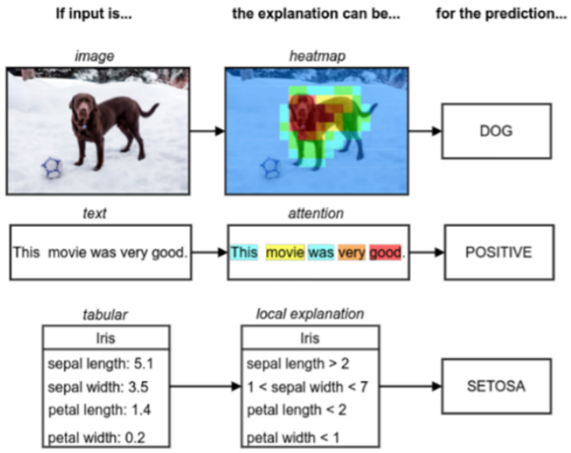



Figure 1.1: Examples of what explanations can look like for different data types, from Ras G. et al. (2020) [5].

be, for example, the ears and eyes of a cat in an animal classifier. For this project, we will be focusing on a more widely beneficial classification problem – differentiating between Malignant and Benign breast tumours. We will be applying 3 existing Explainable AI (XAI) techniques to a pre-existing convolutional neural network (CNN) trained on breast mammograms, and visualising the explanations produced by each method. We will then go on to compare and contrast these methods, and determine whether they agree with each other, and with the opinions of a radiologist, to the point that we could trust them for diagnostic purposes.

## 1.1 Motivation

The recent developments of Machine Learning has sparked an interest in the possibility of its use in more specialised and high stakes fields, for example medical diagnostics. Given a particular medical scan, a clinician may want to be able to differentiate between healthy and unhealthy tissue, or between different types of pathologies. However, the black-box nature of ML models means their conclusions tend not to be trusted by clinicians, who cannot determine how the model came to a particular decision. Interpretable explanations are therefore crucial. In fact, many medical experts have already expressed their concerns over the problems presented by the black-box deep learning approaches which are commonly used as the current state of the art, as discussed in a report by Jia X. et al. (2020) [13]. Because of this, the application of XAI techniques for medical diagnostics is an extremely active field of research.

Using XAI techniques for medical diagnostics is an exciting and important field, however it comes with its own set of problems. Medical datasets are difficult to work with due to differing labelling standards – some images will have complete clinical annotation, while others will simply state whether a tumour is present in the image or not. There is also the issue of gaining access to these datasets, which are usually protected by governing data laws. Explanation techniques also tend to run into problems when faced with images with small Regions of Interest (ROIs). This is because many

XAI techniques first split an image into segments, and may consequently gloss over some very important pixels. This is the case for breast mammograms as the images are large, and cancerous regions can be extremely small. Medical images are also noisy and blurry by nature, leading to potential confusion by classifiers when it comes to masking out unimportant parts of an image.

A review by Rudin C. (2019) [2] discusses the serious implications of bad explanations in high stakes contexts such as medicine and law. Saliency maps, or heatmaps, which are commonly used to visualise image explanations, can be virtually identical for different classes on the same image. Rudin [2] also states that some XAI techniques can simply fail to grasp the model they are applied to, and produce results based solely on similarities between training and test data. Unreliable and misleading explanations are remarkably common, and can have serious negative implications.

These problems are what has motivated this research – by applying three common XAI techniques to medical scans, we aim to discover and discuss their reliability. If the techniques consistently highlight agreeing regions as being the most important, we could conclude with some confidence that these regions contain Malignant or Benign cancerous cells, and therefore that the XAI techniques are reliable. To check this, we will show a small subset of explanations to a radiologist, in order to check the correctness of the outputs with respect to the medical truth. If they do not agree, however, we could conclude that applying these existing XAI techniques to a medical problem as specific as this is not a feasible approach, and that specialised XAI techniques are needed for reliable diagnostic explanations.

## 1.2 Contributions

The major contributions of this project are as follows:

- Adapted a CNN meant for brain tumour detection for use with breast mammograms to determine whether they contained Malignant or Benign tumours. Our optimal model has a test set accuracy of 96.43%.
- Built structured Python code which applies LIME, RISE and SHAP to breast mammograms using our optimal model.
- Statistically compared explanations to determine whether the techniques agree with each other to the point where their decisions could be trusted in a high-stakes environment such as medical diagnostics.
- Discovered that these three existing XAI techniques tend not to agree according to results drawn from One-Way ANOVA, RBO and Kendall's Tau statistical tests. Plausible reasons behind this are discussed in Chapter 6.
- Consulted with a radiologist and discovered that each technique also performs poorly with respect to the medical truth. The results and comments from this are discussed in Section 5.5.

# Chapter 2

# Background

In this section we will briefly introduce the concepts and techniques necessary for understanding this research. This will include common XAI techniques, Machine Learning concepts, and evaluation techniques.

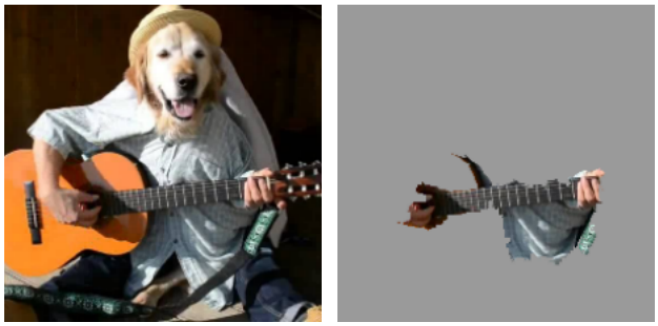
## 2.1 Popular Explainable AI Techniques

Explainable AI is an extremely active research field – new techniques are constantly being published, and the collection of existing techniques is rather large. For the purposes of this section, we will be focusing on describing the three XAI techniques we have used in our experiments, as well as a small number of other popular and interesting techniques. The techniques we have used in our experiments are LIME (Section 2.1.1), RISE (Section 2.1.2) and SHAP (Section 2.1.3). We will be describing the techniques generally, as they appear in their literature – for the three utilised techniques, their code implementations will be described in Chapter 4.

### 2.1.1 LIME - Local Interpretable Model-Agnostic Explanations

LIME, first introduced by Ribeiro M. et al. (2016) [18], is an interpretable XAI technique which can be applied to any model without needing any information about its structure. It is a local technique, which means it explains one particular input – in this case, one breast mammogram.

In practise, LIME provides a local explanation by replacing a complex neural network (NN) locally with something simpler, for example a linear or Ridge regression model. LIME creates many perturbations of the original image by masking out random segments (superpixels), 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 as shown in Figure 2.1, taken from Ribeiro's paper [18].



(a) Original Image (b) Explaining *Electric guitar*

Figure 2.1: An example of an image explanation generated by LIME, from Ribeiro M. et al. (2016) [18].

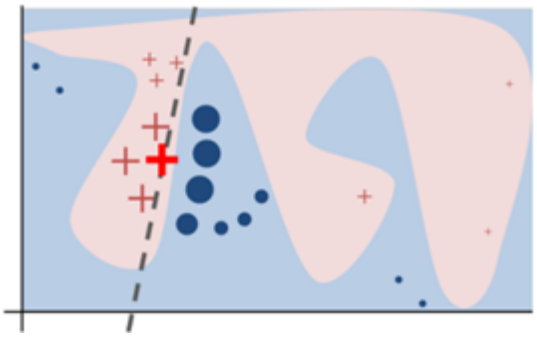


Figure 2.2: How LIME generates local explanations, from Ribeiro M. et al. (2016) [18].

Figure 2.2, also taken from Ribeiro's paper [18] shows us mathematically how LIME forms an explanation. The original model's complex decision function, unknown to LIME, is represented by the blue/pink background boundary. The bold red cross represents the current instance being explained.

LIME samples several perturbations, and obtains predictions using the model's complex decision function. For visualisation purposes, the perturbations are classified as either crosses or circles. It then weighs them by their distance to the explained instance – here the perturbations with higher proximity are shown to be bigger. The dashed line is the learned explanation for the instance, which as you can see is locally faithful, but not necessarily globally faithful, as the model's actual boundary function cannot be described linearly.

### 2.1.2 RISE - Randomized Input Sampling for Explanations of Black Box Models

As introduced by Petsiuk V. et al. (2018) [17], RISE 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 then generated from a linear combination of the masks where the weights come from the output probabilities predicted by the model.


Figure 2.3: Examples of image explanations generated by RISE, from Petsiuk V. et al. (2018) [17].

RISE generates saliency maps which highlight the most important pixels of the image regarding its classification. This makes RISE a very interpretable technique, as the resulting explanation is highly intuitive, as shown in Figure 2.3, taken from Petsiuk's review [17]. RISE is also model agnostic, as it does not require any knowledge of the inner structure of the NN it is using.

It's important to note that this approach is very similar to LIME, however it measures saliency based on individual pixels, rather than superpixels (groups of similar pixels), and therefore may perform better when the region of interest (ROI) of an image is very small with respect to the size of the image itself, for example when identifying tumours in mammograms.

### 2.1.3 SHAP - Shapley Additive Explanations

This model-agnostic approach, introduced by Lundberg SM. (2017) [14], uses Shapley values, a concept from game theory, to find the contribution of each feature to the model's output. Initially, the image is segmented into superpixels, so that we don't need to compute values for each individual pixel. 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. As values decrease, the segments are deemed as less important. Figure 2.4, taken from the SHAP Github repository [20], shows that SHAP is also a highly interpretable method of highlighting the most important segments of an image with regards to its classification. It's important to note here that SHAP values are derived from game theory's Shapley values – they are not the same.

#### 2.1.3.1 SHAP Values

Lundberg's paper [14] highlights the fact that many XAI techniques, such as LIME, use simple explanation models, defined as interpretable approximations of the original model, to explain input classifications. As described in Section 2.1.1, LIME usually utilises a simple regression model or a linear model as its explanation model. Lundberg

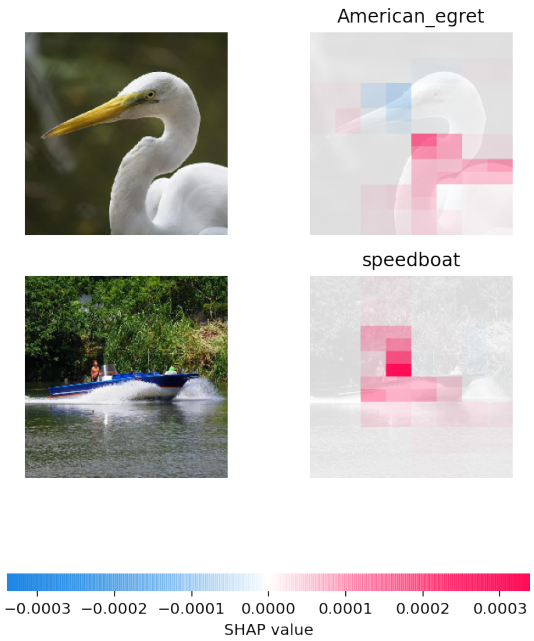



Figure 2.4: Examples of image explanations generated by SHAP, from the SHAP Github repository [20].

[14] also defines Additive Feature Attribution Methods, which are techniques that use explanation models which are linear.

SHAP values are defined as “Shapley values of a conditional expectation function of the original model” [14]. This means that SHAP as a technique is another example of an Additive Feature Attribution Method. For the purposes of this report, details about game theory will be avoided – it is not necessary for understanding of the XAI technique. In short, SHAP values correspond to the change in the expected model prediction for a particular feature, when that feature is binary masked.

### 2.1.4 CAM - Class Activation Mapping

CAM is an example of a model specific technique, whereas the methods described previously have all been model agnostic. This is because CAM requires knowledge of the final layer of the neural network – all fully connected layers at the end of the network must be removed, and replaced with a Global Average Pooling (GAP) layer, which averages over all activation maps to ensure all important parts are caught. This does however limit the usage of this technique to certain neural networks.

This technique is described in detail in a paper by Zhou B. et al. (2016) [10], though in short, the GAP layer outputs the spatial average (the mean over multiple points in space) of the feature map of each unit at the last convolutional layer in the network. Each unit is weighted by the gradient of the output class w.r.t the unit, and these values are summed to generate the final output. This results in a saliency map indicating the parts of the image that activate the class the most – the most important pixels.

### 2.1.5 Grad-CAM - Gradient-Weighted Class Activation Mapping

This method is a generalization of CAM, which does not require a Global Average Pooling layer. It can be applied directly to any Convolutional Neural Network, even those with fully connected layers, and does not require any alterations to the network architecture.

As described in a paper by Selvaraju R.R. et al. (2017) [8], Grad-CAM utilizes the gradients of the model output w.r.t. the final convolutional layer's feature map, to understand the importance of each neuron to the classification decision. Large gradients imply that the model output depends most on these points, making them the most important pixels regarding the classification.

Although CAM and Grad-CAM are both very successful Explainable AI techniques, and have been used to try to explain medical images in the past, they are only applicable to Convolutional Neural Networks (CNNs) with certain structures. This is because convolutional layers tend to retain spatial information that is lost in other types of model layers. The model we are working with in this project, as described in Section 3.3, is convolutional, however its structure does not support these methods – therefore, we opted to use model-agnostic approaches like LIME, RISE and SHAP.

## 2.2 Machine Learning Concepts

Explainable AI techniques require a Machine Learning model to explain. In this section, basic ML concepts such as neural networks, classification problems and simple models such as Ridge regression used in LIME will be described. The focus of this research is not on the ML components, however, so the purpose of this section is purely to provide background.

### 2.2.1 Supervised Learning

Supervised ML algorithms learn based on example input-output pairs, for example a labelled training set of images. They then infer a function based on this data, in order to predict the labels of new unseen data. We will be performing supervised learning using a Convolutional Neural Network (CNN).

Relevant examples of supervised ML algorithms include Neural Networks, Ridge Regression, and Linear Regression.

### 2.2.2 Linear Regression

Linear Regression aims to fit a straight line to a set of data points in a way that models the relationship between the data's parameters as accurately as possible. For example, in a hypothetical 2-dimensional space, linear regression aims to draw the best possible straight line on an X-Y scatter graph of data points.

The best line is defined as the line that produces the lowest MSE (Mean Squared Error) – defined as the average squared residual, where residuals are simply the distances

between the data points and the line.

More formally, the calculation for MSE is shown in Equation 2.1. $y_i$ is the ith observed value, $\hat{y}_i$ is the corresponding predicted value (value on the line), and n is the number of observations.

$$MSE = \frac{(\sum y_i - \hat{y}_i)^2}{n} \tag{2.1}$$

### 2.2.3 Ridge Regression

Ridge Regression, as introduced by McDonald G.C. (2009) [11], is an enhancement of linear regression meant to reduce overfitting and improve robustness. Since Linear Regression chooses the 'best' line based only on MSE values, it commonly chooses lines with very steep gradients when based on a small number of data points – this is a form of overfitting.

Ridge Regression tackles this problem by allowing us to minimise the MSE while also keeping model parameters low – we introduce a penalty function to the parameter values, generating a new loss function for the model to minimise as shown in Equation 2.2.

$$Loss = MSE + \alpha \cdot \sum \theta_i^2 \tag{2.2}$$

Where $\theta_i$ is the ith parameter of the model, and $\alpha$ controls how heavily penalised each model parameter is.

Ridge Regression finds the 'best' line to fit to the data by minimising this new loss function. It is widely preferred to Linear Regression as it does not overfit as dramatically, and this is why it is commonly used as a simple explanation model by XAI techniques such as LIME.

### 2.2.4 Overfitting

One major problem that commonly surfaces when training any kind of Machine Learning model is overfitting – when the model learns a function that is too close to the patterns and noise in the training data, and therefore performs poorly on unseen data. The objective when training any machine learning model should be to maximise performance on the unseen test data, rather than the training data, for this reason.

As shown in a review by Raymond L. et al. (2016) [6] one major source of overfitting is small training sets. This is particularly a problem when considering medical data – gaining access to large datasets is difficult, due to different annotation standards and ethics approvals. Another factor that Raymond highlights as influential to overfitting is large input dimensionality. Since we are working with medical scans, which tend to contain a large number of pixels, this is also extremely relevant.

To try and minimise the effect of overfitting on our results for this project, we will produce models of the same architecture trained with different numbers of epochs (repetitions), and use the model with the highest accuracy on a held-out validation set for producing our explanations. The details of this experiment are in Section 3.4.

### 2.2.5 Classification

Classification is the task of using a ML model to predict a class label for an input, given a set of existing labels in the domain. Classification tasks can be binary, meaning there are two classes to choose from, or multi-class, meaning there are more than two possible classes.

Here, we are considering a binary classification task, where the possible class labels are "Malignant" and "Benign". These predictions are made on single breast mammograms of pixel size 227x227 – the dataset used will be described in detail in Section 3.1.

### 2.2.6 Neural Networks

We will be applying XAI techniques to a complex neural network, which will be described in detail in Section 3.3. Although some XAI methods do approximate the complex network with a simpler regression model, as with LIME in Section 2.1.1, most explanations come directly from the original network.

Neural networks, as described in an introductory review by Krogh A. (2008) [1], are made up of layers of 'neurons', which work as mini threshold functions. The first layer takes the data, for example an image, as an input. The intermediate layers all take the previous layer's output as their input, and the final layer produces the model's prediction. In the case of this specific binary classification, this will be a value between 0 and 1, where values below 0.5 imply Benign, and values over 0.5 imply Malignant.

There are many types of layers, some of which will be described in detail in Section 3.3 alongside the architecture of the Neural Network we will be using.

As mentioned by Selvaraju R.R. (2017) [8], Convolutional Neural Networks are networks that contain convolutional layers – that is, layers that retain spatial information such as proximity. This makes them especially useful for image recognition, as they tend to perform better on pixel data. Deep Neural Networks (DNNs) are neural networks that contain a large number of hidden (intermediate) layers – they are more generally used for deep learning problems, as they can handle more complex functions. The model we will be using for this project is a CNN with many hidden layers - it can be considered as a Deep CNN.

### 2.2.7 Evaluating Model Performance

For this project, we will be evaluating the performance of a CNN with two statistics – Accuracy and F1 Score.

Accuracy is simply the proportion of correct classifications for a given dataset, as shown in Equation 2.3. C is the number of correct classifications, and N is the number of

images in the dataset upon which the model is being evaluated.

$$Accuracy = \frac{C}{N} \tag{2.3}$$

F1 Score is described in Equation 2.4, where TP stands for True Positives, FP stands for False Positives, and FN stands for False Negatives.

$$F_1 = \frac{TP}{TP + 0.5(FP + FN)} \tag{2.4}$$

In this context, True Positives are Malignant images that are correctly classified as Malignant. False Positives are Benign images incorrectly classified as Malignant. False Negatives are Malignant images incorrectly classified as Benign. A perfect model would have a F1 Score of 1.

### 2.2.8 Ground Truth

Ground Truth, otherwise known as Gold Standard. refers to the proven true label applied to a data point. In this context, as the dataset contains images already labelled as Benign or Malignant, we take these labels to be the Ground Truth.

## 2.3 Breast Tumours

The major focus of this project is using XAI techniques to classify breast mammograms into two categories – Benign and Malignant. The dataset, as will be described in Section 3.1, gives us a large collection of labelled images, separated into these two categories. It's important to note that in the real world, the difference may not be so distinct, and may be influenced by other patient factors. For the scope of this project, we are simply focusing on the distinction between these tumour types as laid out in the labelled dataset, and whether Explainable AI techniques can be reliably used to explain our model's predictions.

A recent paper by Reid A. et al. (2019) [7] goes into significant detail about different classes of breast legions, and though this level of detail is out of scope, we feel it is important to briefly outline the differences between Malignant and Benign lesions here.

Reid [7] explains that Benign tumours may present as an "observable breast deformity" such as cysts, rubbery regions, or other textural irregularities. By contrast, Malignant tumours, though significantly more dangerous due to their ability to spread, are more commonly identified by imaging than by dermatologist examination.

This brings to light the importance of imaging, and consequently the importance of a correct conclusion being drawn from a breast mammogram, whether by a clinician or a CNN. Our goal here is to determine whether existing and widely used XAI techniques perform well enough to hold up to these high-stakes applications in the real world.

## 2.4 Statistical Comparison Techniques

After applying LIME, RISE and SHAP to our CNN, we will have generated an explanation for each method, for each image in our test set. For a given image, looking at these three explanations side by side is not enough to definitively decide whether they support each other – formal statistical techniques are required.

### 2.4.1 One-Way ANOVA

As described by Ross A. (2017) [19], a One-Way ANOVA "compares the means of two or more groups for one dependent variable". Groups of a size greater than 30 are required to prevent false conclusions, and ANOVA compares the variation within these groups to the variation based on the group means.

To make this test feasible, for each image in our test set we will be separating out the 'most important' pixels according to each of the three explanation methods. To determine how many pixels should be in these lists, we will empirically choose a number L (Section 4.1.3) of 'most important features' for LIME, and use the top n pixels for RISE and SHAP, where n is defined as the number of pixels within the top L LIME features. This is because we want to compare the similarities between pixel lists of equal lengths between methods.

The need for the difference in method between LIME and the other techniques here comes from the differing ways of weighting pixel importance. LIME outputs a binary value for each pixel - whether they appear in the L most important features or not. RISE and SHAP however allocate decimal importance values to pixels.

After these pixel lists are generated, we will simply calculate their percentage agreement, defined here as the proportion of pixels these lists have in common between methods. The input groups for the One-Way ANOVA experiment will therefore be three lists, of length equal to the test set size, which will be greater than 30. These will be lists of percentage agreements between methods, labelled LIME-RISE, LIME-SHAP, and RISE-SHAP. Percentage agreement in this context will be treated as the dependent variable for the ANOVA test. The fixed (independent) variable will be the XAI techniques being compared.

One-Way ANOVA uses the following hypotheses:

- H0 (Null Hypothesis): 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.

When applied to two or more groups, One-Way ANOVA outputs a F-statistic and a p-value. If the p-value is higher than the alpha value, we can say that there is no statistically significant difference between the means of the groups. If the p-value is lower than the alpha value, we can say that the difference is statistically significant, and reject the null hypothesis.

The results of this experiment will highlight any inconsistencies in our explanations by

determining how similar each pairwise XAI technique comparison is, with respect to their top n pixel agreements. Alpha values will be defined in Chapter 5.

### 2.4.2 Kendall's Tau

First introduced by Kendall M. (1938) [15], Kendall's Tau is a non-parametric measure of the degree of the correlation between two ranked lists. The method focuses on the ranks of items within a list, rather than their numerical values, and can therefore be applied to ordered categorical variables, for example pixel IDs of the form "(x, y)".

Taking the lists of all image pixels and their numerical importance values for each of the 3 methods, we will first rank them in descending order of importance value, generating three ordered lists of pixel IDs – one for each XAI technique. The mathematical approach to calculating Tau values is as follows, as taken from Kendall M. (1938) [15]:

According to basic probability, for any sample of n paired observations, there are $m = \frac{n(n-1)}{2}$ possible comparisons of the points $(X_i, Y_i)$ and $(X_j, Y_j)$. Observations are concordant if $X_j - X_i$ and $Y_j - Y_i$ have the same sign, and if they do not, the observations are defined as discordant. If we take C to be the number of concordant pairs, and D to be the number of discordant pairs, then Tau is defined as in Equation 2.5.

$$Tau = \frac{C - D}{m} \tag{2.5}$$

A large value of $(C-D)$ indicates a strong positive relationship between the two ordered lists, as there is a predominance of concordant pairs. If $(C-D)$ is a large negative value, this indicates a strong negative relationship between X and Y. Here, $m$ acts as a normalizing coefficient so that Tau values are always in the range [-1, 1].

The purpose of this test is to discover whether two ordered lists are independent, as would be implied by a Tau value of 0. Similarly to One-Way ANOVA, a significance test is used, with the following hypotheses:

- H0 (Null Hypothesis): There is no statistically significant correlation in pixel ordering, the lists are independent.
- H1: There is a statistically significant correlation in pixel orderings between the lists, they are not independent.

A review by Abdi H. (2007) [12] includes a table of critical Tau values to be used in hypothesis testing for lists of maximum size 10. For longer lists, Abdi [12] states that null hypothesis tests can instead be performed with a Z value which is normally distributed with mean 0 and standard deviation 1. This value is shown in Equation 2.6.

$$Z = 3 \cdot Tau \cdot \frac{\sqrt{n(n-1)}}{\sqrt{2(2n+5)}} \tag{2.6}$$

From this Z value, any online Z-score to p-value calculator such as the one provided by Statology [21] can be used to find the corresponding p-value. Existing code implemen-

tations, like the Python package we will be using, generate p-values automatically. For p-values higher than the alpha value, the result is not statistically significant, and we cannot reject the null hypothesis – the lists are independent. For p-values smaller than the alpha value, we can reject the null hypothesis, and say that there is a statistically significant correlation between the lists.

Applying this test to the ordered pixel ID lists from each of the three Explainable AI techniques will allow us to determine whether there is a significant correlation between them. We would expect each technique to rank pixels similarly in order to generate agreeing explanations, however different techniques rank pixels in different ways, so discussing how the results of this test are affected by different techniques will be interesting and relevant to this research. These differences will be discussed in Section 5.3.

### 2.4.3 RBO - Ranked Biased Overlap

Kendall's Tau, though useful, has one major drawback in this context – it applies equal weightings to pixels no matter where they are in the list. Concordant or discordant pixel pairs at the bottom of the list, with the lowest pixel importance, will be weighed equally to pairs at the top of the list, with the highest pixel importance. We are trying to determine whether XAI techniques highlight the same pixels as being the most important – therefore, we want a comparison technique that applies more weight to the pairs at the top of the lists.

Introduced by Webber W. (2010) [24], RBO solves this problem by using weights for each rank position – it considers the depth of the ranking being examined. This has the desired effect of minimising the effect the 'tail' of the lists has on the final result. In this context the 'tail' would be the pixels considered most unimportant by the XAI technique – for example, background pixels.

Taking two ranked lists as inputs, RBO outputs a single value between 0 and 1, where 0 indicates that the lists are completely disjoint, and 1 indicates that they are identical. This value is calculated as in Equation 2.7, which is taken directly from Webber W. (2010) [24].

$$RBO(S,T,p) = (1-p) \cdot \sum_{d=1}^{\infty} (p^{d-1} \cdot A_d) \tag{2.7}$$

Where S and T are two ordered lists, p is a tuneable parameter, d is the position of the element in the list, and $A_d$ is the proportion of pixels in agreement at depth d.

There is no hypothesis testing for this method. The output value will simply tell us how similar the ranked pixel ID lists for each Explainable AI method are – we will be experimenting with this in Section 5.4.

# Chapter 3

# Model Setup

This chapter focuses on introducing the dataset and CNN we will be using for our Explainable AI experiments. We will detail all image pre-processing steps undertaken, and include model performance analysis on the validation set.

## 3.1 The Original Dataset

For this project we are taking breast mammograms with cancerous masses from a public dataset [22], which takes images from three official datasets – INbreast [4], DDSM [3] and MIAS [9]. When generating this dataset, the creators 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 all images to 227x227 pixels.

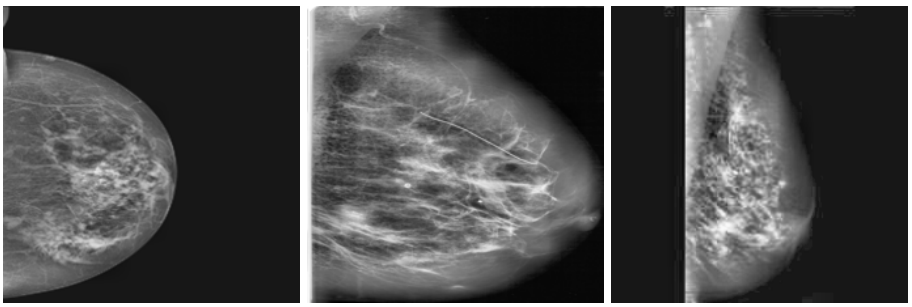
## 3.2 Data Pre-Processing

### 3.2.1 Image Selection

The public dataset [22] we are using is rather large, with 7632 INbreast images, 3816 MIAS images, and 13128 DDSM images after data augmentation. Attempting to run our CNN on these raw datasets led to memory problems, so we elected to manually cut down the dataset to a more manageable level.

The original paper that introduced this dataset, by Huang M. et al. (2020) [23], details their data augmentation techniques, which includes rotating and flipping each image to generate 14 variations of itself. For the purposes of this research, this is not useful – 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. The first pre-processing step was therefore manually selecting all scans of the original orientation, which were helpfully labelled within the dataset.

The second decision made was to only take images from INbreast and DDSM. The only MIAS scans present in the dataset were Benign, and since images from each of the 3



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

Figure 3.1: Example images from each of the three dataset sources.

datasets have visual differences, it's important to have images from each source for both classification decisions. The visual difference in original scans between the three sources is illustrated in Figure 3.1.

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.

### 3.2.2 Image Cropping

The first step of altering these selected images for maximal model performance was to crop out as much of the black background as possible, making the breast the focus of each image. This was performed using basic Python libraries such as opencv, and an example outcome is shown in Figure 3.2. After cropping, the images are resized to the original 227x227 pixel format for consistency.

### 3.2.3 Splitting the Dataset

In order to train a CNN on this new dataset, it first needs to be split into Training, Validation and Testing sets. The model will be trained on the Training set. The Validation set will be used for all intermediate experiments – deciding how many epochs to train the model for (Section 3.4), and tuning parameters for LIME (Section 4.1.2). The Testing set will be used only once, when generating the final explanations using the optimised XAI techniques – this set will be used for statistical analysis and final project outcomes. Our dataset of 2236 images is split into a (Training / Validation / Testing) ratio of (2124 / 56 / 56).

## 3.3 Model Architecture

For this project we are using an existing public CNN [16] which was originally used for binary classification of brain scans into categories 'Yes' and 'No', regarding the presence of a tumour. In the original study the model achieved 88.7% accuracy on a test set of brain scans.

To make this model compatible with our data, we had to change the input size from 240x240 pixels to 227x227 pixels to accommodate for the slightly smaller size of the

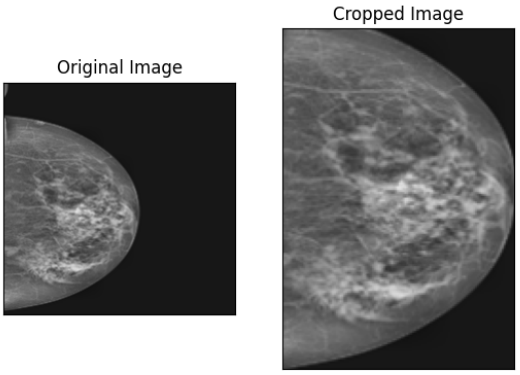



Figure 3.2: Example of image cropping.

breast mammograms. Otherwise, the architecture is identical. 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 classifications, so any output value over 0.5 means the model has classified the image as Malignant.

This CNN contains 8 layers:

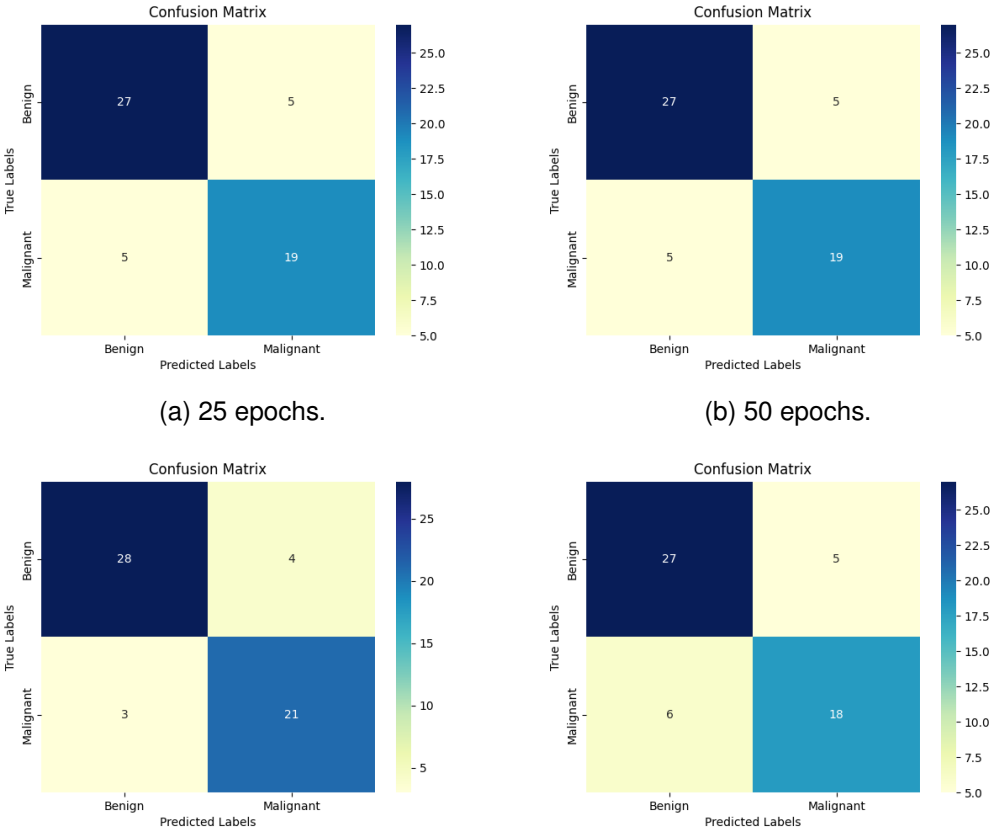
1. Zero Padding Layer – Applies 2 pixels with value 0 to each side of the image.
2. Convolutional Layer (Section 3.5.1)
3. Batch Normalization Layer – Normalizes input matrix to maintain a mean close to 0 and a standard deviation close to 1.
4. ReLU Activation Layer – Converts any negative values to 0. Returns matrix with positive values only.
5. Max Pooling Layer (Section 3.5.2)
6. Max Pooling Layer (Section 3.5.2)
7. Flatten Layer – Flattens 3D image matrix into 1-dimensional vector.
8. Dense (Fully Connected) Layer – Outputs one value between 0 and 1.

### 3.3.1 Convolutional Layer

This layer slides a 7x7 kernel over the image matrix, moving one pixel at a time, and performs elementwise multiplication. The result is that each 7x7 section of the image is converted into one value, decreasing the size of the matrix.

### 3.3.2 Max Pooling Layer

This layer reduces the dimensionality of images by sliding a 4x4 kernel over the image matrix, moving it one pixel at a time. For each 4x4 section, it returns the maximum




(a) 25 epochs.

(b) 50 epochs.

(c) 75 epochs.

(d) 100 epochs.

Figure 3.3: Confusion Matrices corresponding to model performances on the Validation set, for differing numbers of epochs.

pixel value, resulting in a smaller matrix which is still representative of the regions of the original image.

## 3.4 Training the Model

When training any neural network, we want to run the training set through the model more than once, to ensure that the model has the chance to correctly learn the function of the data. The number of repetitions is commonly described as the number of epochs. As discussed in Section 2.2.4, a major issue when training any neural network is overfitting. If the number of epochs is too high, we risk overtraining the model, and reducing its accuracy on any unseen data. If the number of epochs is too low, we risk underfitting the data, which is the opposite problem – the model would not be able to capture enough of the trends within the dataset, and would be too general.

To ensure that we are generating explanations on a reliable model, we have trained four models, and will evaluate their performances on the Validation set. These models all have the architecture described in Section 3.3, and are trained on the Training set described in Section 3.2.3, however they are trained with different numbers of epochs.

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

Table 3.1: Table of Validation Accuracies and F1 Scores for our CNN trained with different numbers of epochs.

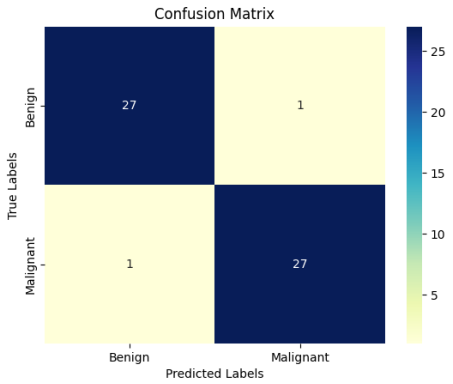



Figure 3.4: Test set Confusion Matrix for CNN trained on 75 epochs.

The performances of these models are described in Table 3.1, where the Accuracy and F1 Score (Section 2.2.7) are shown. The confusion matrices for the models are shown in Figure 3.3. These statistics are based on a Validation set of 56 images.

The 75 epoch model performs the best, and will therefore be the model we proceed with. This is to be expected, as in general accuracy will increase with number of epochs until the point of overfitting is reached. Here, the 100 epoch model has started to overfit. The 75 epoch model's performance on the Test set is described in Table 3.2 and Figure 3.4.

## 3.5 Importance of Explanations

Even though the accuracy of our model on the Test set has a high value of 96.43%, the model is not perfect and does make 2 classification errors. In high-stakes contexts such as medical diagnosis, the instances where the model is incorrect could be extremely costly. In addition, real-world unseen data may be unlabelled. We would need to be able to implicitly trust that the model's output is correct. This is why it is important that we understand exactly how the model comes to its conclusions – the purpose of XAI.

| Test Accuracy | Test F1 Score |
|---|---|
| 0.9643 | 0.9642 |

Table 3.2: Table of Test Accuracy and F1 Score for our CNN trained on 75 epochs.

# Chapter 4

# Explanations

In this chapter, we will showcase the individual explanations generated by each of the 3 Explainable AI techniques, when applied to our CNN and the Test set of 56 images. We will show 6 Benign and 6 Malignant examples for each technique, with more being displayed in the Appendices for optional viewing. For LIME, empirical experiments regarding parameter choices were undertaken – the results of which will also be presented here. For each technique we will display explanations for the same 12 images, and discuss any cases where the techniques perform poorly.

## 4.1 LIME Explanations

### 4.1.1 Methodology

The code for generating LIME explanations follows these steps for each image in the dataset it is applied to:

1. Segment the image into superpixels using built-in Python segmentation functions. This involved empirically choosing parameters for the segmentation algorithm, the process for which is explained in Section 4.1.2.
2. Generate perturbations of the image as described in Section 2.1.1, and predict their labels using our CNN.
3. Weight these perturbations by their closeness to the original image, using an exponential kernel of width 25.
4. Fit a Ridge Regression model (Section 2.2.3) to the data, using the weights gathered for each perturbation to fit the local model (the CNN).
5. Separate out the L most important features, and then generate the explanation using our Python code. The process for choosing how many features to use is detailed in Section 4.1.3.
6. Showcase the explanation by highlighting the boundaries of the L most important features in yellow – this is the best way to show the features without obscuring the image.

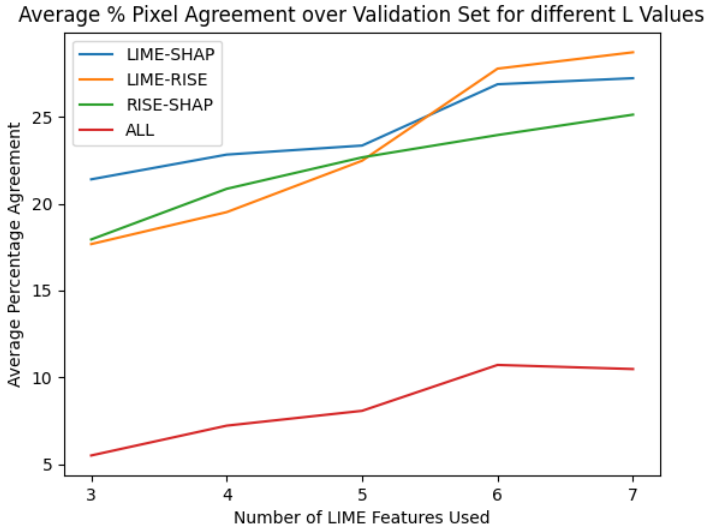



Figure 4.1: Average pixel agreement taken over 30 images from the Validation set. Comparing each two methods separately, and then comparing all three.

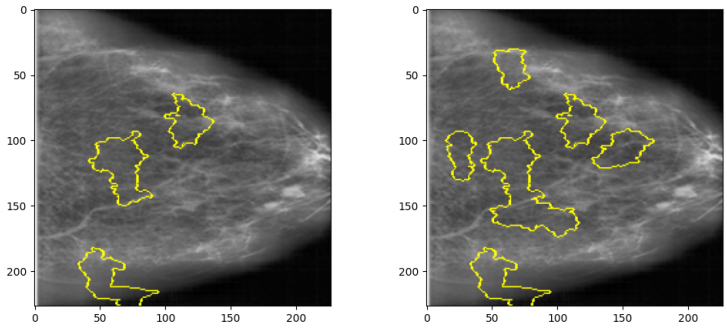
### 4.1.2 Choosing Segmentation Parameters

The segmentation algorithm used by LIME here is the built-in scikit-image quickshift algorithm, which has three tuneable parameters – the kernel size, the max-dist value, and the ratio value.

Kernel size dictates the size of the kernel which we use to slide across the image in order to identify superpixels. The smaller the kernel size, the more superpixels the image will be separated into. Since we are working with medical scans with small regions of interest, we want to avoid large superpixels as too much information about the image would be lost. Therefore, small kernel sizes are preferred – we use a value of 2 for this experiment.

Max-dist controls the cut-off point for data distances, and higher values for this parameter lead to fewer superpixels being generated. Since we want the image to be segmented into a large number of small superpixels, we want a low value for this parameter. Here we use a value of 10, which is the default value.

The ratio value determines how much weight is given to colour-space and image-space proximities. Large values mean more weight is given to the colour-space, meaning the similarities between pixel colour values. Smaller values give more weight to the proximities of pixels to each other. For this value we experimented with both high and low ratios, and discovered that for high ratios, the most important superpixels tended to include background regions. Therefore, we are using a low ratio value of 0.1 for these experiments – we want more weight to be given to the proximities of the pixels, rather than their colours, to avoid the pixels near the boundary of the breast being deemed as incorrectly important.



(a) L = 3 (b) L = 7

Figure 4.2: Difference between LIME Explanations using the top 3 most important features (a), and the top 7 (b).

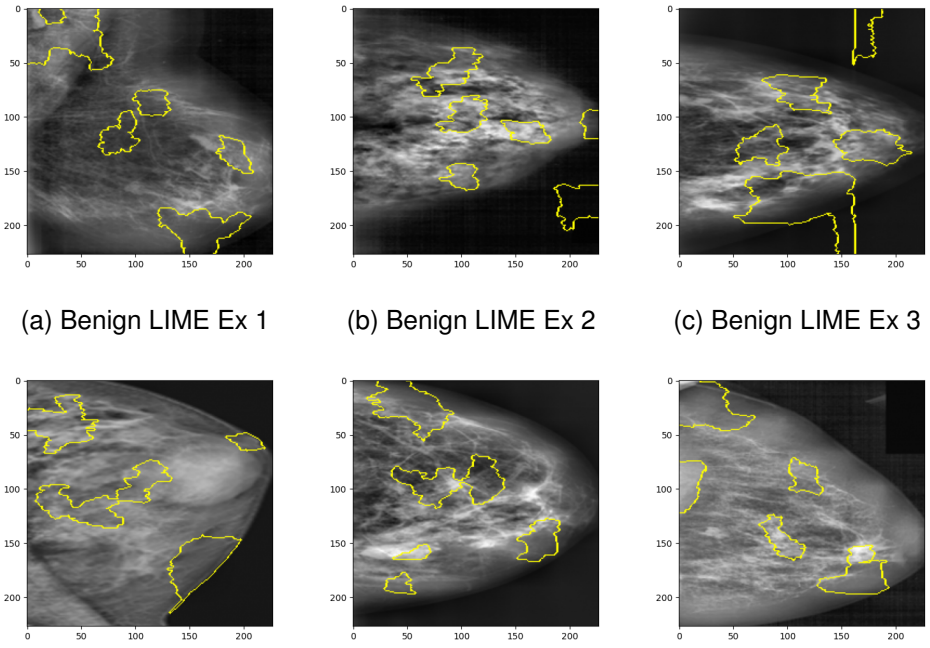
### 4.1.3 Choosing L

We define L as the number of most important features (superpixels) used in explanations – this is a tuneable parameter. The L value we choose will determine the features shown in the individual LIME explanations, and will also determine how many pixels are compared in the One-Way ANOVA statistical comparisons (Section 5.2).

As described in Section 2.4.1, the process for comparing the lists of most important pixels between each XAI technique is important to our understanding of how much the techniques agree with each other. The lists compared will contain the same number of pixels for each technique – the number of pixels in the L most important LIME features. Therefore, to determine which L value to use, we calculated the average percentage pixel agreements between methods for lists with differing lengths – using L values of 3, 4, 5, 6 and 7. Averages were taken over the first 30 images in the Validation set, for the sake of time. The results of this experiment are shown in Figure 4.1.

In general, as L increases, average pixel agreement increases between each pairwise technique comparison. This is consistent with what we would expect – as more pixels are considered, it is more likely for the lists to have more pixels in common. However, we can't simply keep increasing L, as we are trying to compare the most important pixels, and as we consider more pixels we are comparing pixels of lower importance.

For this reason, we have chosen L to be 6. This is because 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. It is also interesting to note here that the increase for the RISE-SHAP comparison seems to be constant for L values between 4 and 7. The difference in explanations generated by LIME for L values of 3 and 7 is shown in Figure 4.2.



(a) Benign LIME Ex 1 (b) Benign LIME Ex 2 (c) Benign LIME Ex 3

(d) Benign LIME Ex 4 (e) Benign LIME Ex 5 (f) Benign LIME Ex 6

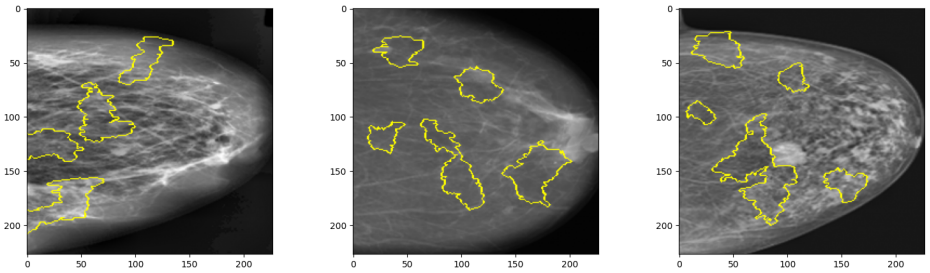
Figure 4.3: Examples of LIME explanations generated for Benign breast mammograms.

### 4.1.4 Explanation Examples

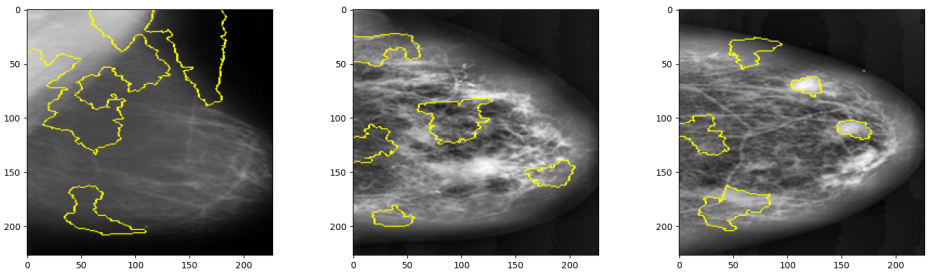
Figure 4.3 shows six examples of LIME explanations generated for images whose true labels are Benign. Figure 4.4 shows six LIME examples for Malignant scans. More examples of LIME explanations can be found in Appendix A.

For both classes, some explanations highlight undesirable features such as the image background. These inconsistencies are likely due to the variance in breast shape throughout the dataset, which can clearly be seen in these examples. This is unavoidable, though could be lessened by using larger datasets in the future.

Looking at these explanations without the medical knowledge required to visually identify cancerous regions tells us little about whether these are 'good' explanations – whether they are highlighting regions which are actually cancerous. To evaluate LIME's performance, we will compare these outputs to those of RISE and SHAP, in order to determine whether the techniques support each other enough to be trusted in this high-stakes context. We will also cross-reference a small subset of explanations with the opinions of a radiologist, to determine whether the explanations generated by each technique agree with the medical truth.



(a) Malignant LIME Ex 1 (b) Malignant LIME Ex 2 (c) Malignant LIME Ex 3



(d) Malignant LIME Ex 4 (e) Malignant LIME Ex 5 (f) Malignant LIME Ex 6

Figure 4.4: Examples of LIME explanations generated for Malignant breast mammograms.

## 4.2 RISE Explanations

### 4.2.1 Methodology

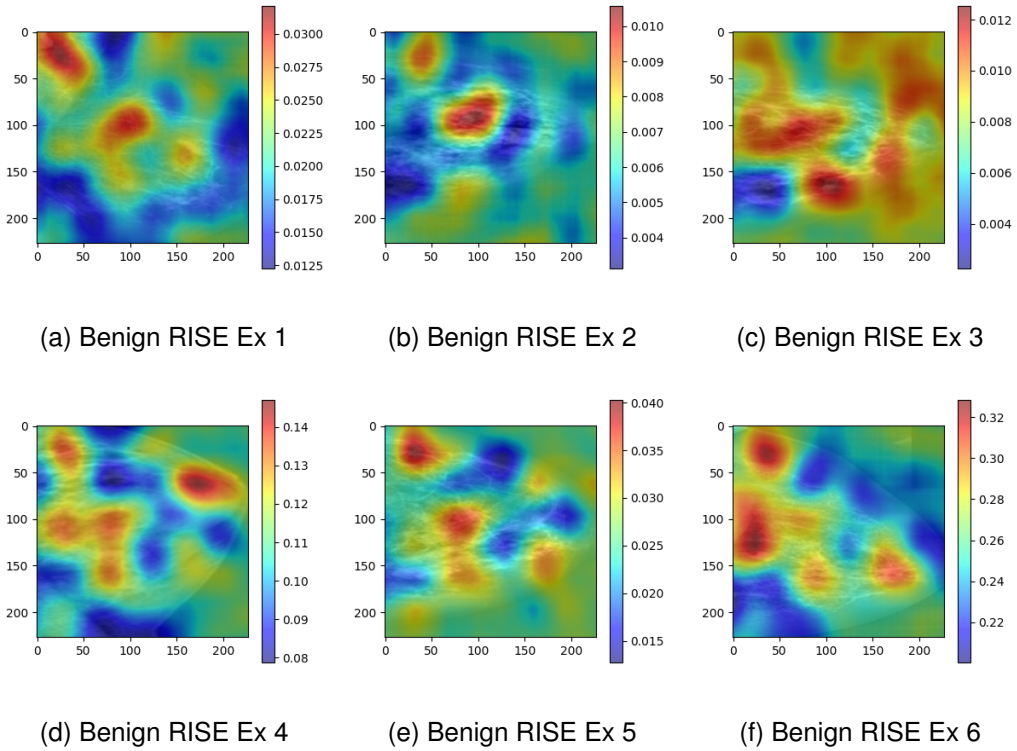
The code for generating RISE explanations follows these steps for each image in the dataset it is applied to:

1. Generate 2000 random masks of the same size as the input image.
2. Perform elementwise multiplication of each of these masks with the input image, and run each of these outputs through our CNN to get their predictions.
3. Generate saliencies of pixels using these masks and their predictions, as described in Section 2.1.2.
4. Visualise these pixel saliencies as a heatmap, showcased over the original image.

There were no tuneable parameters for this method.

### 4.2.2 Explanation Examples

Figure 4.5 shows six examples of RISE explanations generated for images whose true labels are Benign. Figure 4.6 shows six RISE examples for Malignant scans. More examples of RISE explanations can be found in Appendix B. In these heatmaps, the most important pixels are shown as red, and the least important pixels are shown as blue. It's important to note that RISE considers values of individual pixels, whereas



(a) Benign RISE Ex 1 (b) Benign RISE Ex 2 (c) Benign RISE Ex 3

(d) Benign RISE Ex 4 (e) Benign RISE Ex 5 (f) Benign RISE Ex 6

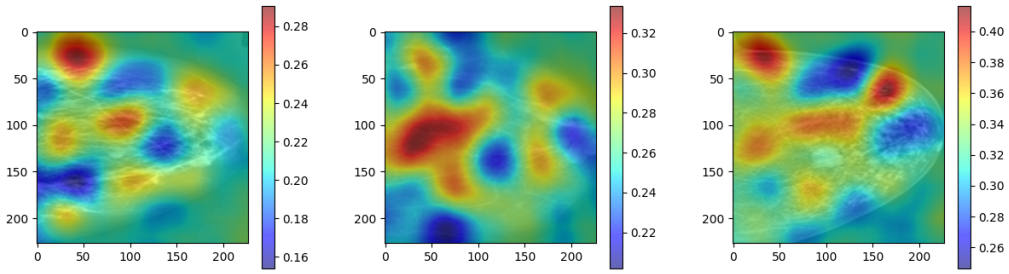
Figure 4.5: Examples of RISE explanations generated for Benign breast mammograms.

LIME and SHAP consider superpixels. It's also important to note that each image has a different importance value scale.

### 4.2.3 Observations

For both classes, RISE does not disregard the background of the images as much as it ideally should - backgrounds are generally shown with yellow/green, signifying a medium relative importance. Like LIME, we expect that this is due to the differing breast shapes, as it seems to be a consistent issue.

LIME and RISE seem to generate poor results for the same images - we define poor results here as explanations which highlight background regions as incorrectly important. In particular, Figure 4.6 (d) shows a case where RISE performs especially poorly for a Malignant image. Interestingly, LIME also performs poorly on this image, as shown in Figure 4.4 (d). This image does have an irregular shape, which supports our thoughts. Another example is the image shown in Figure 4.3 (c) and Figure 4.5 (c) - a Benign image for which the background is deemed as incorrectly important for classification by both LIME and RISE.



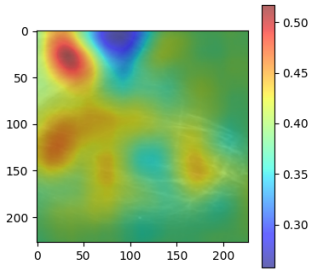
(a) Malignant RISE Ex 1



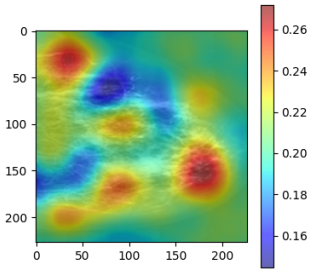
(b) Malignant RISE Ex 2



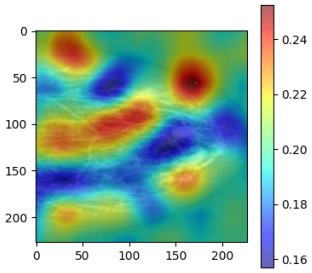
(c) Malignant RISE Ex 3



(d) Malignant RISE Ex 4



(e) Malignant RISE Ex 5



(f) Malignant RISE Ex 6

Figure 4.6: Examples of RISE explanations generated for Malignant breast mammograms.

## 4.3 SHAP Explanations

### 4.3.1 Methodology

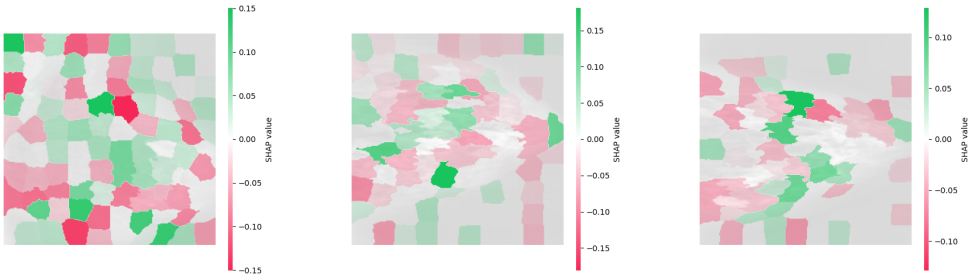
The code for generating SHAP explanations follows these steps for each image in the dataset it is applied to:

1. The image is segmented into 100 segments (the default value), and masks are generated where segments are randomly hidden.
2. The Kernel Explainer from the SHAP Python API is used to explain the CNN. The explainer uses the CNN's predictions of the masked images.
3. SHAP values are generated for each segment using this explainer.
4. Each segment is assigned a colour based on its SHAP value, and the explanation is displayed as these coloured regions overlayed over the original image.

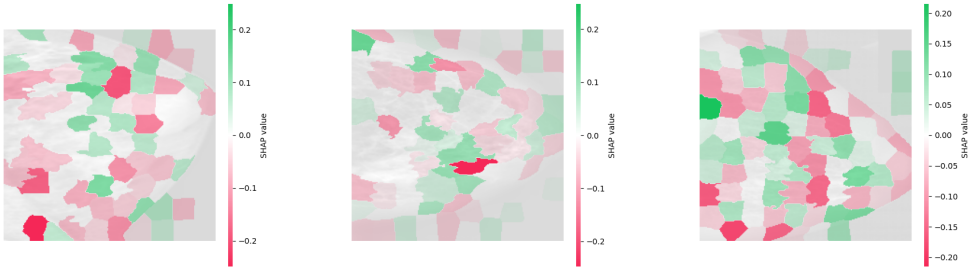
There were no tuneable parameters for this method.

### 4.3.2 Explanation Examples

As with LIME and RISE, we show six examples of SHAP explanations for Benign images in Figure 4.7, and six examples of SHAP explanations for Malignant images in Figure 4.8. More examples of SHAP explanations can be found in Appendix C.

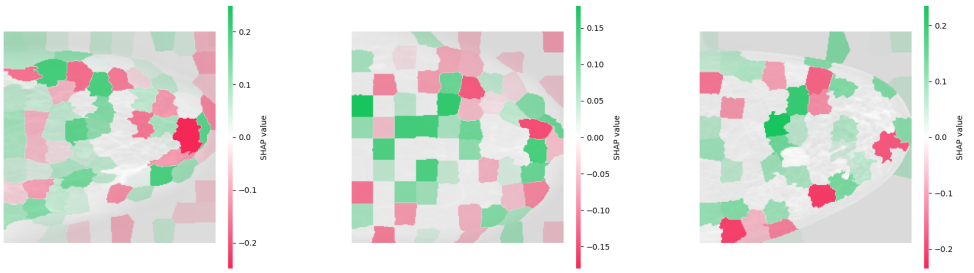



(a) Benign SHAP Ex 1 (b) Benign SHAP Ex 2 (c) Benign SHAP Ex 3

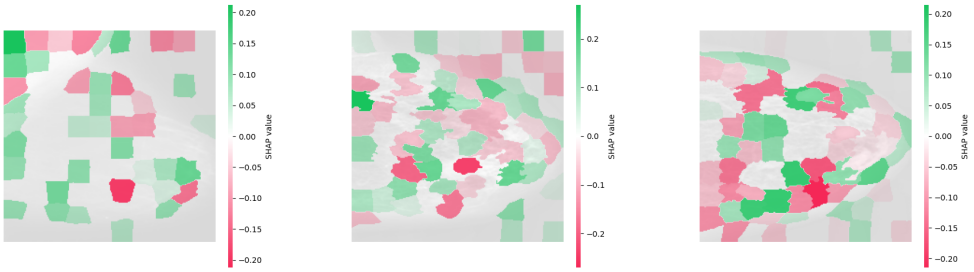



(d) Benign SHAP Ex 4 (e) Benign SHAP Ex 5 (f) Benign SHAP Ex 6

Figure 4.7: Examples of SHAP explanations generated for Benign breast mammograms.

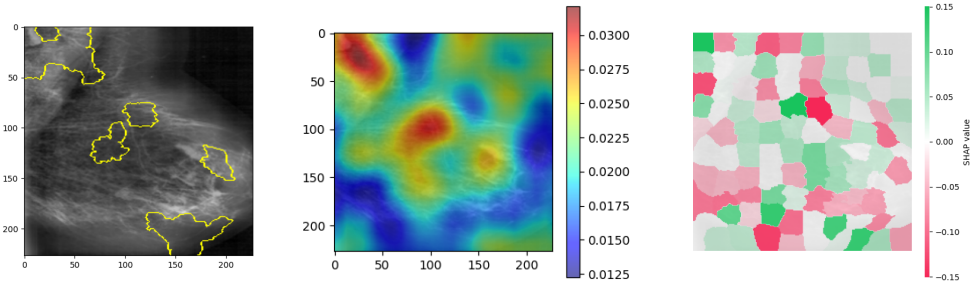



(a) Malignant SHAP Ex 1 (b) Malignant SHAP Ex 2 (c) Malignant SHAP Ex 3




(d) Malignant SHAP Ex 4 (e) Malignant SHAP Ex 5 (f) Malignant SHAP Ex 6

Figure 4.8: Examples of SHAP explanations generated for Malignant breast mammograms.




(a) LIME Explanation

(b) RISE Explanation

(c) SHAP Explanation

Figure 4.9: Explanations generated by all 3 XAI techniques for an image.

Segments that contribute the most (high SHAP values) to the classification of the image are shown as green. The least important segments are assigned negative SHAP values, and are shown as red. It's important to note that the SHAP value scales are not the same for all images.

### 4.3.3 Observations

As previously discussed for the other XAI techniques, the fourth Malignant image has an irregular breast shape, and therefore LIME and RISE performed poorly. SHAP is no exception, as shown in Figure 4.8 (d). There is a large number of heavily influential segments at the top-left corner of the image, indicating that the irregular breast shape has again had an effect on the explanation.

In most cases, the superpixels outside the boundary of the breast seem to have lower SHAP values, meaning they are less influential to the model's prediction. This is not always the case, but looking at these examples, SHAP seems to disregard background pixels with more success than RISE.

## 4.4 Need For Statistical Analysis

Looking at the three explanations side-by-side for an image, as in Figure 4.9, we can start to see some agreement. In this example, regions highlighted as red by RISE and green by SHAP seem to be roughly the same as the ones highlighted by LIME. However, due to the different ways of visualising explanations between each technique, we cannot determine how much they agree with each other by simply looking at them. Statistical comparison methods are necessary - these experiments will be detailed in Chapter 5.

# Chapter 5

# Evaluating Explanations

In this Chapter we will explore the levels of agreement between explanations generated by LIME, RISE and SHAP. We will start by visualising the pixels deemed as most important (using the L value determined in Section 4.1.3) in an overlaid plot, and then go on to complete multiple statistical analyses on the raw pixel importance data from each technique. The radiologist evaluation will be detailed at the end of this Chapter.

## 5.1 Visualising Agreement

In Section 4.1.3, we empirically decided to use the top 6 most important features in our LIME explanations. For the remainder of this chapter, we denote n to be the number of pixels within these features. For comparison methods such as One-Way ANOVA (Section 2.4.1), we want to compare pixel agreement within these lists of n pixels, where pixel agreement is defined simply as the proportion of pixels the method's lists have in common. Therefore, we first visualise the n most important pixels according to each technique, and critique whether they seem to overlap. Figure 5.1 shows some examples of these plots. The n most important pixels as decided by SHAP are shown in blue, with RISE in red and LIME in green. More examples of these plots can be found in Appendix D.

Figure 5.1 shows that there are some points of overlap, where all 3 techniques have identified the same regions as highly important. However, there are many regions in which the techniques disagree. Figures (b) and (d) from Figure 5.1 show cases where explanations have performed poorly - defined here as identifying background pixels to be the most important. As previously discussed in Chapter 4, this is likely due to irregular breast shapes within the dataset. Figures 5.1 (a) and (c) show instances where explanations have multiple clear points of agreement.

Although these plots do intuitively show the agreements and inconsistencies between techniques for their n most important pixels, they are informal representations of agreement. The rest of this chapter will explore more formal statistical analyses.

| 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 |

Table 5.1: Results of One-Way ANOVA tests performed on pairwise method comparisons of average pixel agreement lists for top n most important pixels. **Bold** results are statistically significant.

## 5.2 One-Way ANOVA Results

As described in Section 2.4.1, we want to compare the average pixel agreements of the XAI techniques, where pixel agreement is defined as the proportion of pixels two lists have in common. We use lists of length n (where n is the number of pixels in the top 6 LIME features for an image) for each of the three techniques, and compute a percentage value to represent the pixel agreement between each pair of techniques.

The groups for this ANOVA test are LIME-RISE, LIME-SHAP and RISE-SHAP. For each group, the input is a list of percentage pixel agreements for each image in the test set. The dependent variable will be pixel agreement, while the fixed variable will be the XAI techniques being compared. We will set the alpha value as 0.05.

The hypotheses for One-Way ANOVA are as follows:

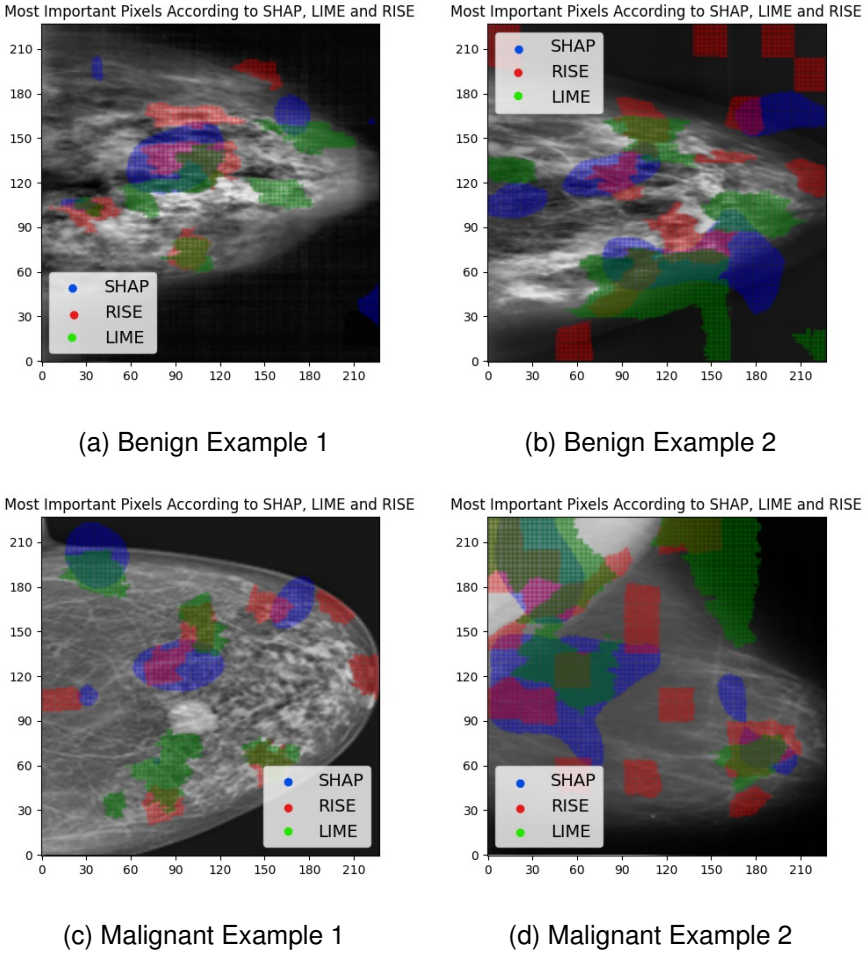
- H0 (Null Hypothesis): There is no statistically significant difference between the means of the groups.
- H1: There is a statistically significant difference between the means of the groups.

We perform this test using the built-in Python method from the scipy package. The results are shown in Table 5.1.

For Tests 1 and 3, the p-values generated are greater than the alpha value, so we can say that there is no statistically significant difference between the means of the groups compared. Test 1 tells us that LIME has no statistically different agreement level when compared to RISE or SHAP. Test 3 tells us that SHAP has no statistically different agreement level when compared to LIME or RISE. Test 2 however has a p-value lower than the alpha value, meaning we reject the null hypothesis, and that there is a statistically significant difference between the means of the groups. This tells us that RISE has a statistically significant difference in agreement level when compared to LIME or SHAP.

We visualise these results in a boxplot as in Figure 5.2, where medians are represented by orange lines and means are represented by green triangles. The largest difference in mean value comes between LIME-RISE and RISE-SHAP, which supports the conclusions of this One-Way ANOVA test. For this plot, we have also included pixel agreement statistics from comparing all 3 methods together, for visualisation purposes.

As shown in both Figure 5.2 and Table 5.2, the average pixel agreement between techniques is startlingly low – 20 - 30% on pairwise comparisons, and under 10% when




(a) Benign Example 1

(b) Benign Example 2

(c) Malignant Example 1

(d) Malignant Example 2

Figure 5.1: Plots of overlap between the 3 XAI techniques regarding the n most important pixels. Figures (a) and (b) are Benign, while Figures (c) and (d) are Malignant. SHAP is shown in blue, RISE in red, and LIME in green.

| Techniques | Mean | Standard Deviation | 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% |

Table 5.2: Statistical overview of percentage pixel agreements for all method comparisons.

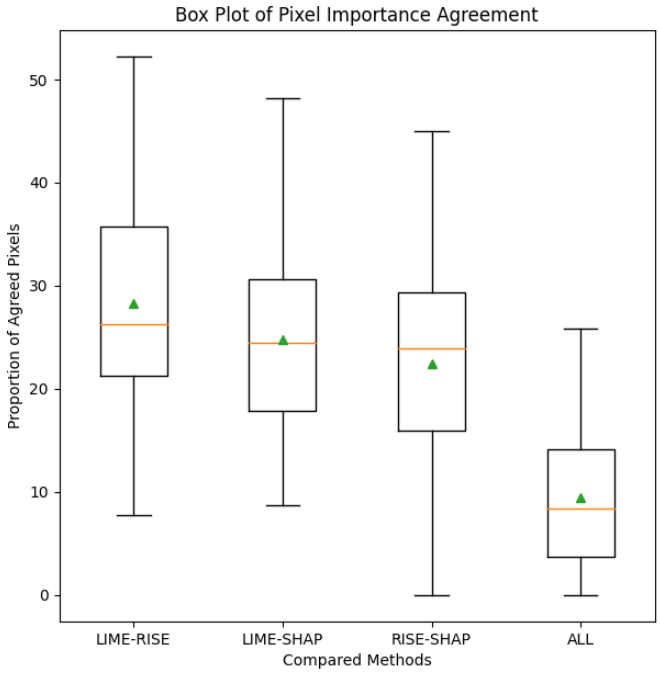



Figure 5.2: Boxplot of Pixel Agreement statistics between XAI techniques for top n most important pixels. Medians are shown as orange lines, and means are shown as green triangles.

comparing all three. However, it is important to note that there is significant standard deviation across all images in each list, and also that these mean pixel agreements still represent significant number of pixels, as our images are large and have small regions of interest. This will be discussed further in Chapter 6.

## 5.3 Kendall's Tau Results

As described in Section 2.4.2, the purpose of Kendall's Tau is to discover whether two ordered lists are independent. We perform this test using the built-in Python method from the scipy package, and set the inputs to be the ordered lists of pixels and their importance values for each of the three XAI techniques. These lists are in descending order, and each item is of the form "(x, y): value".

The following hypotheses are used:

- H0 (Null Hypothesis): There is no statistically significant correlation in pixel ordering, the lists are independent.
- H1: There is a statistically significant correlation in pixel orderings between the lists, they are not independent.

We applied Kendall's Tau to each image in the test set 3 times – on the full pixel list, on a list of the n most important pixels, and on the 1000 most important pixels. This

| Techniques | p-values | | | Tau | | |
|---|---|---|---|---|---|---|
| | Full | n | 1000 | Full | n | 1000 |
| RISE-SHAP | 0.1227 | 0.1254 | 0.2490 | 0.0032 | 0.0018 | 0.0008 |
| LIME-SHAP | **0.0000** | **0.0481** | 0.0672 | 0.1543 | 0.1056 | 0.2927 |
| LIME-RISE | 0.0662 | 0.0552 | 0.1328 | 0.0041 | -0.0059 | 0.0139 |

Table 5.3: Kendall's Tau results performed on ordered pixel importance lists of varying lengths for each technique. The definition of n here is the same as is stated at the beginning of Section 5.1. Values are averages taken over the test set, and shown to 4 decimal places. **Bold** results are statistically significant.

is because we want to determine not only how much the techniques are agreeing in general, but also whether they are agreeing on the pixels that are deemed most important to the classification, and therefore the diagnosis.

For each pixel list comparison, we report the average p-values and Tau values in Table 5.3 – these averages are taken over the test set. P-values smaller than the alpha value tell us that there is a statistically significant correlation between pixel orderings, and that the lists are not independent. Positive Tau values tell us that the correlation is positive, while negative Tau values tell us that it is negative - the closer these values are to 0, the smaller the correlation.

Setting alpha to be 0.05 (as with ANOVA), we can draw the following conclusions:

- All three lists for RISE-SHAP are independent.
- Both the full pixel list and the list of the n most important pixels for LIME-SHAP show statistically significant positive correlation, as shown by average p-values smaller than 0.05, and positive average Tau values. The top-1000 pixel list is independent.
- All three lists for LIME-RISE are independent. However, we note that the average p-values are much closer to the threshold of 0.05 and would be considered statistically significant if a larger alpha value was used.

These conclusions imply that LIME and SHAP have the highest correlation, while RISE and SHAP have the lowest. Kendall's Tau gives equal weightings to agreeing pairs regardless of where they are in the list – a problem discussed in Section 2.4.3, and addressed by RBO. Before we go on to perform RBO analysis however, there are some interesting comparisons to note between the results of Kendall's Tau and One-Way ANOVA.

ANOVA compares pixel agreement, defined as the proportion of pixels two lists have in common, regardless of their order. In Figure 5.2, we see that for the top n pixel lists, LIME and RISE have the highest mean pixel agreements, while RISE and SHAP have the lowest. Kendall's Tau has also highlighted RISE and SHAP to be the methods with the lowest correlation in pixel orderings. However, it has highlighted LIME and SHAP to be the techniques with the most correlation. This implies that while LIME and RISE have higher pixel agreement regarding the presence of the same pixels in the top n pixel

| - | 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 |

Table 5.4: 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. Raw values can be found in Excel files associated with this project.

lists, LIME and SHAP agree the most regarding pixel order.

## 5.4 RBO Results

Ranked Biased Overlap (RBO), as described in Section 2.4.3, weights each pair of pixels by its position in the ordered list. This makes it a more useful test in this context – we want to apply more weight to the pixels at the top of the list, as we care more about agreement with regards to the pixels which were actually used in a diagnosis.

RBO experiments output single values between 1 and 0, where 1 implies the lists are identical, and 0 implies they are entirely disjoint. The results of RBO depend on the tuneable parameter p – this value can be thought of as the probability that a hypothetical manual analyser will continue on to the next pair in a list, instead of stopping. Smaller p values therefore place more weight on the items at the top of an ordered list.

While we do want to apply more weight to pixels at the top of the ordered pixel lists, we have to keep in mind the differences in pixel importance value allocation between XAI techniques. RISE operates on single pixels, applying an individual importance score to each pixel. SHAP uses superpixels, meaning each pixel within a given segment will be assigned the same importance value. LIME in this case has a binary importance scoring system – simply whether the pixels are in the top 6 most important features or not. For this reason, we must use large p values to ensure that we are considering enough of the list to properly encompass similarities between larger groups of pixels with identical values.

Table 5.4 shows the average, minimum and maximum RBO values for each pairwise pixel list comparison. Looking solely at the average RBO values for each pairwise comparison, this test tells us that the ordered pixel lists are almost disjoint. However, this is not unexpected, due to the differing pixel importance value allocations between techniques.

What is interesting about these results are the maximum values – LIME and SHAP generate lists that are hugely identical for at least one instance in the test set, generating maximum RBO values in the range 0.69 – 0.78. The other pairwise comparisons do not come close to these numbers - this will be discussed further in Chapter 6. This observation supports the results from Kendall's Tau – both tests have identified LIME

| 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 | 2 | 2 | 2 | 2 | 2 | 1 | 2 |
| SHAP | 0 | 0 | 0 | 1 | 2 | 0 | 1 | 1 | 1 | 0 |

Table 5.5: Radiologist evaluation regarding explanations generated on a subset of 10 images. B denotes Benign, and M denotes Malignant. The meanings of the assigned values are explained in Section 5.5.

and SHAP as the techniques with the highest agreement regarding pixel orderings.

## 5.5 Radiologist Evaluation

The previous analysis within this Chapter has focused on determining whether our 3 XAI techniques agree with each other. It is also important to understand whether the regions being highlighted by each technique are actually cancerous – for this, we have consulted a radiologist, and provided them with a small subset of 10 images, along with the corresponding explanations as generated by each of the 3 techniques. We were unable to gather an expert evaluation for the entire test set due to time constraints and the radiologist's busy schedule.

We asked for results in the form of a numerical value between 0 and 3 assigned to each explanation, which represents the quality of the explanation with respect to the medical truth. The definitions of each number as provided to the radiologist 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

The results we received are shown in Table 5.5. It is interesting to first note that no explanations earned a label of 3 – each explanation either identified erroneous regions or missed important sections of the image. With respect to the medical truth, LIME appears to perform the worst, with 6 instances of quality 0 and 4 instances of quality 1 within this subset of 10 images. SHAP performs slightly better, with 5 instances of quality 0, and one instance of quality 2. It is RISE which performs significantly better than the other methods – it only has one instance of quality 0, with the majority of explanations being of quality 2. This is likely because RISE is the only method which uses pixels rather than superpixels – it is less likely to gloss over small regions of interest within the images. There does not seem to be any huge difference in explanation quality between Benign and Malignant images for any technique.

The radiologist also made some interesting comments which are relevant to this research:

- It is difficult to evaluate the explainability of benign mammograms if there is no single focus legion.

- Explanations only select some image regions and don't label all critical structures.

Note that these 10 results may not be representative of the entire test set. However, the issues highlighted by the above comments are consistent problems.

## 5.6 Threats to Validity

This research was carried out on a relatively small dataset of breast mammograms - we assume that this dataset is representative of the real-world population. We also limit our classification task to Benign or Malignant - there are many external factors that affect a real-world diagnosis, including repeat illness, and there are also many types of lesion for both classes, which would appear differently in our mammograms - we do not consider this. In future work, using a non-binary classification domain alongside a more thorough radiologist evaluation may allow us to better analyse the failures of our techniques.

All empirical analysis regarding LIME parameter tuning, as well as when deciding the number of features to showcase and compare, was based solely on the patterns within our data. They may not hold up when compared to a larger dataset.

We also note that our XAI techniques by definition utilise randomization when generating masks, and as a consequence of this re-running our code will generate slightly different results to the ones displayed in this report. Early experiments with this showed that this variation is not hugely impactful as we generally discuss average values in our statistical tests - the observations we discuss in Chapter 6 generally seem to hold.

Other threats to validity include our CNN, which was taken from a public project on brain tumour detection, and so the network layer parameters were designed for that particular domain. Though the model does perform very well in our context, as shown in Section 3.4, perhaps changing the architecture of our model would yield improved explanations.

We also note that our code for LIME, RISE and SHAP is not the only way of implementing these techniques - there are many public examples of projects involving these explanation techniques online, and each one approaches the problems in a slightly different way while following the same general steps as described in their literature. Because of this, another researcher's code may yield slightly different results to the ones shown here.

# Chapter 6

# Observations and Discussion

In this chapter we will sum up the implications of the results generated in Chapters 4 and 5, and discuss the final contributions of this project. We will also briefly highlight some opportunities for future work in this area.

## 6.1 Observations

### 6.1.1 Observation 1

**Each technique performs poorly on the same images.**

The explanations in Chapter 4 highlight the variation in explanation quality within the test set. Interestingly, each technique performed poorly (highlighted background pixels as most important) on the same images - usually images with irregular breast shapes. This is likely due to the small size of the dataset, and the effect of blurring and image re-sizing as described in Section 3.2.2. It is however interesting that these problems don't seem to impede the model's accuracy, only the quality of the explanations.

### 6.1.2 Observation 2

**RISE and SHAP are the only methods with instances of 0 common pixels within their top n pixel lists.**

Figure 5.2 shows the statistical structure of the lists of pixel agreements regarding the top n pixels – lists of length 56, the size of the test set. Over these 56 images, LIME always has pixel agreement with both other methods - the minimum pixel agreement values are 7.74% for LIME-RISE lists, and 8.73% for LIME-SHAP lists. RISE and SHAP however have instances where they do not agree at all – where their lists of n most important pixels have zero pixels in common.

### 6.1.3 Observation 3

**One-Way ANOVA implies that LIME and RISE have the highest pixel agreement (common pixels) for the top n pixels, while RISE and SHAP have the lowest.**

Using these same pixel agreement lists, the One-Way ANOVA test (Section 5.2) told us that the only pairwise comparison with a statistically significant difference in means was the comparison of LIME-RISE to RISE-SHAP. This implies that RISE has a statistically significant difference in agreement level when compared to LIME or SHAP. We can see from Figure 5.2 that the LIME-RISE comparison has the highest mean of all pairwise comparisons, while RISE-SHAP has the lowest. We can therefore conclude that solely regarding pixel agreement, defined as the proportion of common pixels, LIME and RISE have the highest agreement, and LIME and SHAP have the lowest.

### 6.1.4 Observation 4

**Pixel agreement percentages between each pair of XAI techniques are extremely low. With respect to medical truth, RISE performs the best while LIME performs the worst - however all explanations have errors.**

Although LIME and RISE do appear to have the most pixel agreement, these statistics are still low, with an average percentage agreement of 28.27%. However, this can still represent a significant region of the image, as seen in the overlay plots in Figure 5.1 – 28% of the top n pixels could still be a thousand pixels, and in medical images such as these, with large sizes and potentially small regions of interest, these overlapping regions could be enough for a diagnosis.

The radiologist evaluation presented in Section 5.5 show that each explanation generated tends to miss critical structures within the mammograms. 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. On a subset of 10 images, 6 LIME explanations and 5 SHAP explanations were deemed as entirely in disagreement with the medical truth. RISE performed the best in this context, with 6 explanations labelled as mostly correct. However, no explanations generated for any technique were labelled as perfect.

### 6.1.5 Observation 5

**Kendall's Tau implies that LIME and SHAP have the most similar pixel orderings.**

Our Kendall's Tau analysis (Section 5.3) tells us that regarding pixel orderings, the LIME-SHAP comparison was the only one in which any statistically significant correlation was present – both in the full (227x227) pixel list, and the top-n pixel list. The test also tells us that although all pixel lists are independent for both the RISE-SHAP and LIME-RISE comparisons, it is the RISE-SHAP comparison which yields the worst results. The average p-values for the LIME-RISE comparisons are significantly closer to the alpha value of 0.05, and if we used a slightly higher alpha value of 0.07, both the full pixel list and top-n pixel list would have yielded statistically significant results. This is in agreement with Observations 2 and 3 – RISE and SHAP are the XAI methods which consistently agree the least, both in top n pixel agreement and pixel orderings.

Combining our Kendall's Tau results with those from the One-Way ANOVA test, we can make an interesting conclusion. The LIME-RISE comparison yields the highest average pixel agreement for the top n pixels, where pixel agreement is defined as the

proportion of pixels two lists have in common. However, the LIME-SHAP comparison shows the most correlation in pixel orderings. Therefore, our Kendall's Tau results imply that while LIME and RISE consistently highlight the highest proportion of the same pixels as being the n most important, LIME and SHAP actually have the most similar pixel orderings.

### 6.1.6 Observation 6

**RBO implies that LIME and SHAP have the most similar pixel orderings.**

In Section 5.4, we briefly discussed the problem with analysing pixel orderings in this context - each XAI technique has a very different way of assigning values to pixels. This will be addressed further in Section 6.2. With this in mind, the goal of our RBO analysis was simply to see if there were any instances in the test set where full ordered pixel lists were not disjoint between techniques. The average RBO values of below 0.1 were to be expected – we are analysing order similarity over lists of length 51529 (227x227), where pixels are assigned values in wildly different ways between techniques.

Looking at the maximum RBO values, the LIME-SHAP comparison yielded numbers in the range 0.69-0.79, indicating that for this specific instance the ordered pixel lists were significantly more alike than they were different. The other two comparisons did not come close to these numbers, however it is important to note that the maximum values obtained for the LIME-RISE comparison were consistently higher than those for the RISE-SHAP comparison. This ties directly into Observation 5 – both RBO and Kendall's Tau suggest that when considering the overall ordering of pixels, LIME and SHAP have the highest agreement.

## 6.2 Discussion

The goal of this project was to determine whether taking off-shelf Explainable AI techniques and applying them to a specific medical context was a feasible approach that could generate diagnostic explanations which would hold up in the real world. Bringing together our observations tells us that this is definitely not the case.

Though out of the three, LIME and SHAP are shown to have the highest agreement in pixel orderings, these agreement levels are still remarkably low. Explanations generated by these techniques highlight some of the same areas, though have more disagreements than overlap, and are therefore unreliable for use in diagnostics. The most likely reason that LIME and SHAP are the techniques with the highest pixel ordering agreement is the fact 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. RISE assigns decimal values to each pixel, while SHAP assigns the same decimal value to all pixels within a superpixel. Our LIME code outputs binary importance values – a value of 1 if the pixel is within one of the 6 most important features, and a value of 0 if it is not. This binary scoring method is likely to be the reason behind the slightly higher top-n pixel agreement statistics for pairwise comparisons involving LIME when compared to RISE-

SHAP. These differences come from both the underlying properties of each technique, and from our specific code architecture. These differences could be addressed, and accounted for, in further work on analysing the disparities between the techniques – this would not, however, change the performance of the techniques themselves.

It's important to note that each XAI technique we have applied works very differently, and that the resulting explanations depend on many different factors – how or if it segments images into superpixels, the effect of mask randomization, and the choice of tuneable parameters. While this is an expected reason for some variation in results, a higher level of cohesion in explanations was to be expected – plots such as those in Appendix D should still show significantly more overlapping regions, and lists of percentage pixel agreements as detailed in Table 5.2 should have higher average values. We identified early on 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 relatively small size of the dataset, and the quality of the images 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 XAI techniques showed no noticeable difference in explanation quality for images from the Benign or Malignant classes.

Regarding the medical truth according to a radiologist, explanations generated by LIME and SHAP are commonly deemed as entirely incorrect based on the results discussed for 10 images in Section 5.5. RISE seems to produce the most diagnostically correct explanations, likely due to the fact that it does not use image segmentation, however no explanations were labelled as perfect as areas seem to always be 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.

Thus the final contribution of this project is a message for the active Explainable AI community – LIME, RISE and SHAP do not perform reliably in the context of breast tumour classification. The results of these explanation techniques do not match what a pathologist would require in a real-world context. Instead of pixels or superpixels, techniques should identify and highlight clinically defined regions such as lesions. This is a gap which needs to be bridged - this project highlights the need for specific, carefully defined techniques generated for medical domains, and invites the prospect of future analysis of technique reliability on medical image data.

## 6.3 Opportunities for Future Work

There are many ways in which this work could be further explored. For example:

- Investigating agreement of other XAI techniques, perhaps performing separate analyses on methods that use pixels and superpixels.
- Editing LIME, RISE and SHAP to produce more diagnostically relevant explanations. Segmentation approaches which highlight clinically defined regions such as legions instead of general superpixels could be explored.
- Further consulting with pathologists to analyse failures and determine how to improve these techniques for real-world use.

# Bibliography


[1] Krogh A. What are artificial neural networks? In *Nat Biotechnol 26*, pages 195 – 197, 2008.

[2] Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. In *Nat Mach Intell 1*, pages 206 – 215, 2019.

[3] Heath M. et al. The digital database for screening mammography. In *Proceedings of the Fifth International Workshop on Digital Mammography*, pages 212 – 218, 2001.

[4] Moreira IC. et al. INbreast: toward a full-field digital mammographic database. In *Acad Radiol, 19(2)*, pages 236 – 248, 2012.

[5] Ras G. et al. Explainable deep learning: A field guide for the uninitiated. In *arXiv, 2004.14545*, 2020.

[6] Raymond L. et al. Overfitting in linear feature extraction for classification of high-dimensional image data. In *Pattern Recognition, Volume 53*, pages 73 – 86, 2016.

[7] Reid A. et al. Skin diseases of the breast and nipple: Benign and malignant tumours. In *Journal of the American Academy of Dermatology, Volume 80, Issue 6*, pages 1467 – 1481, 2019.

[8] Selvaraju R.R. et al. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *arXiv, 1610.02391*, 2017.

[9] Suckling J. et al. Mammographic image analysis society (MIAS) database v1.21. In *https://www.repository.cam.ac.uk/handle/1810/250394*, 2015.

[10] Zhou B. et al. Learning deep features for discriminative localization. In *arXiv, 1512.04150*, 2016.

[11] McDonald G.C. Ridge regression. In *WIREs Comp Stat, 1*, pages 93 – 100, 2009.

[12] Abdi H. Multiple correlation coefficient. In *Salkind N.J., Ed., Encyclopedia of measurement and statistics*, pages 648 – 651, 2007.

[13] Cai L. Jia X., Ren L. Clinical implementation of AI techniques will require interpretable AI models. In *Med. Phys. 47*, pages 1 – 4, 2020.

[14] Lee S. Lundberg SM. A unified approach to interpreting model predictions. In *Proceedings of the 31st international conference on neural information processing systems*, pages 4768 – 4777, 2017.

[15] Kendall M. A new measure of rank correlation. In *Biometrika 30*, pages 81 – 89, 1938.

[16] MohamedAliHabib. Brain tumour detection, Github repository. In *https://github.com/MohamedAliHabib/Brain-Tumor-Detection*. GitHub, 2019.

[17] Saenko K. Petsiuk V., Das A. RISE: Randomized input sampling for explanation of black-box models. In *arXiv, 1806.07421*, 2018.

[18] Guestrin C. Ribeiro M., Singh S. "Why should I trust you?": Explaining the predictions of any classifier. In *arXiv, 1602.04938v3*, 2016.

[19] Amanda Ross and Victor L. Willson. *One-Way Anova*, pages 21 – 24. SensePublishers, Rotterdam, 2017.

[20] slundberg. shap, Github repository. In *https://github.com/slundberg/shap*. Github, 2022.

[21] Statology. Z-score to p-value calculator. https://www.statology.org/z-score-to-p-value-calculator/, 2018. Accessed: 04-03-2022.

[22] M Huang. T Lin. Dataset of breast mammography images with masses. In *Mendeley Data, V5*, 2020.

[23] M Huang. T Lin. Dataset of breast mammography images with masses. In *Data in Brief, Volume 31, 105928*, 2020.

[24] Zobel J. Webber W., Moffat A. A similarity measure for indefinite rankings. In *ACM Transactions on information systems*, volume 28, 4, 2010.

# Appendix A

# Further LIME Examples

This section contains further examples of LIME explanations generated for images in the test set. As you can see, some explanations seem feasible, while others are clearly not - for example those which identify background pixels as important.

Explanations highlight the boundaries of the 6 most important features in yellow. 6 is an empirically chosen value as discussed in Chapter 4 of this report.

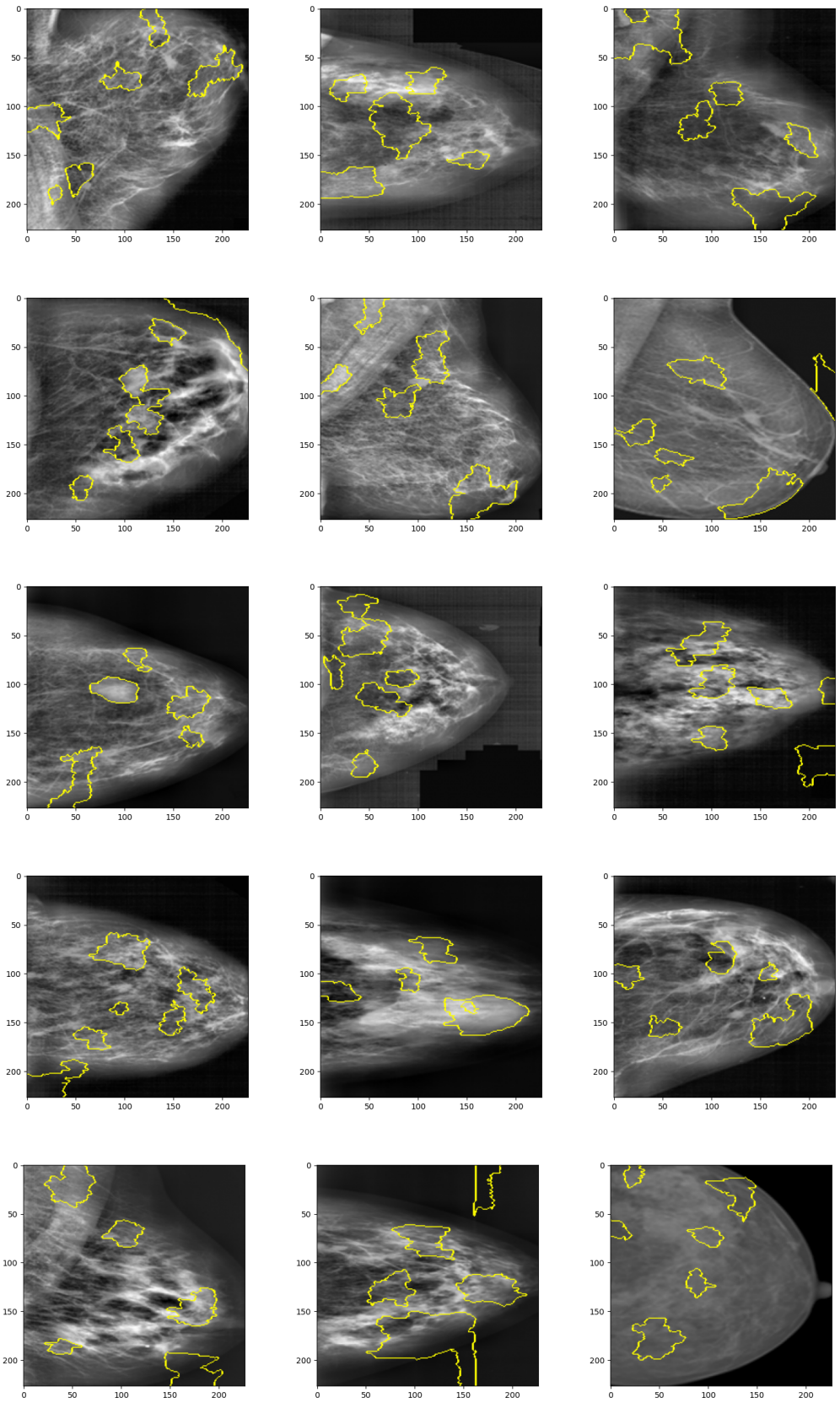


Figure A.1: Further Benign LIME Explanations

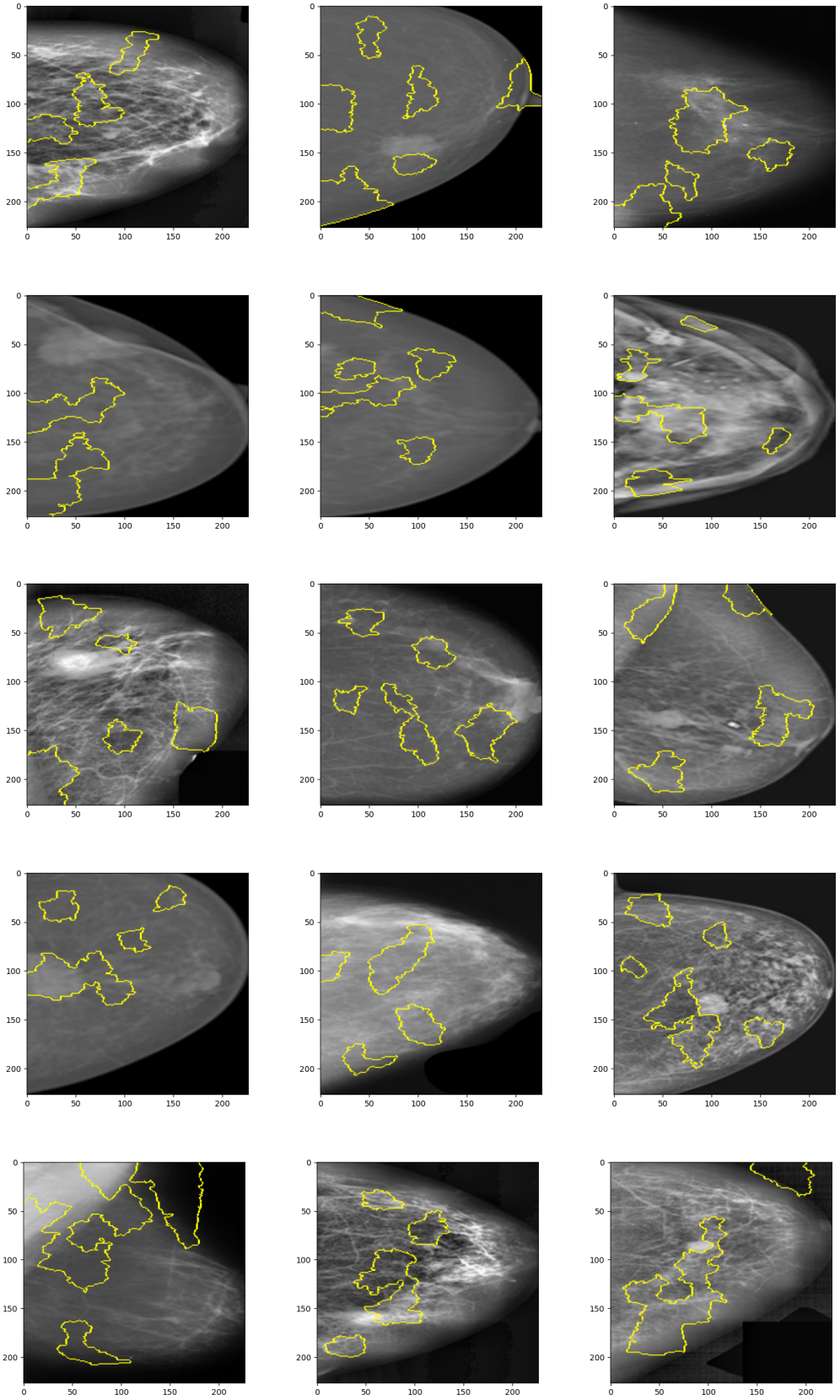


Figure A.2: Further Malignant LIME Explanations

# Appendix B

# Further RISE Examples

This section contains further examples of RISE explanations generated for images in the test set.

It is important to note that the scales of each saliency map are not the same - different magnitudes of importance values are calculated for different images. Highly important pixels are shown in red, with less important pixels shown in blue.

RISE is less adept at disregarding the backgrounds of images, as clearly seen by these examples.

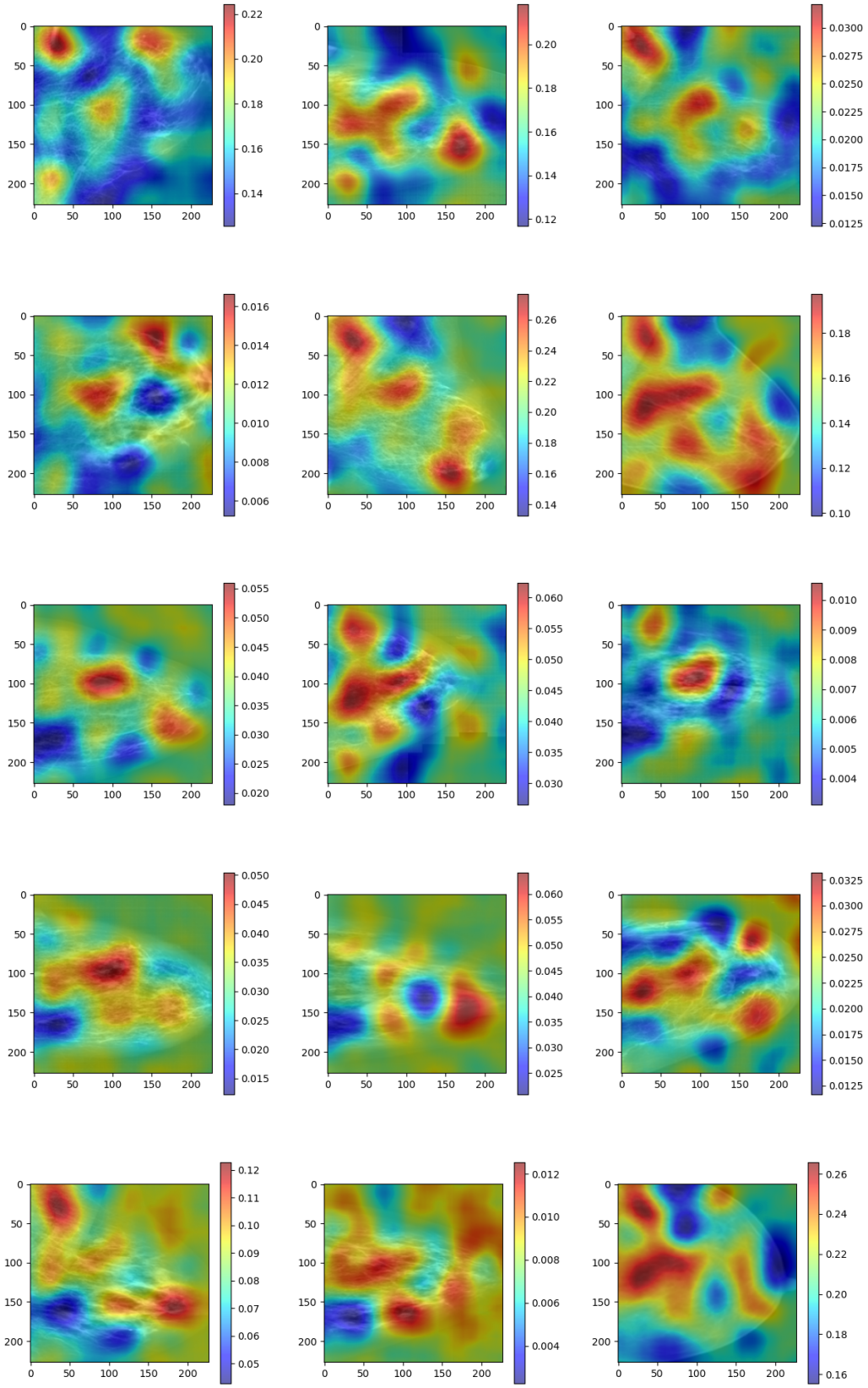


Figure B.1: Further Benign RISE Explanations

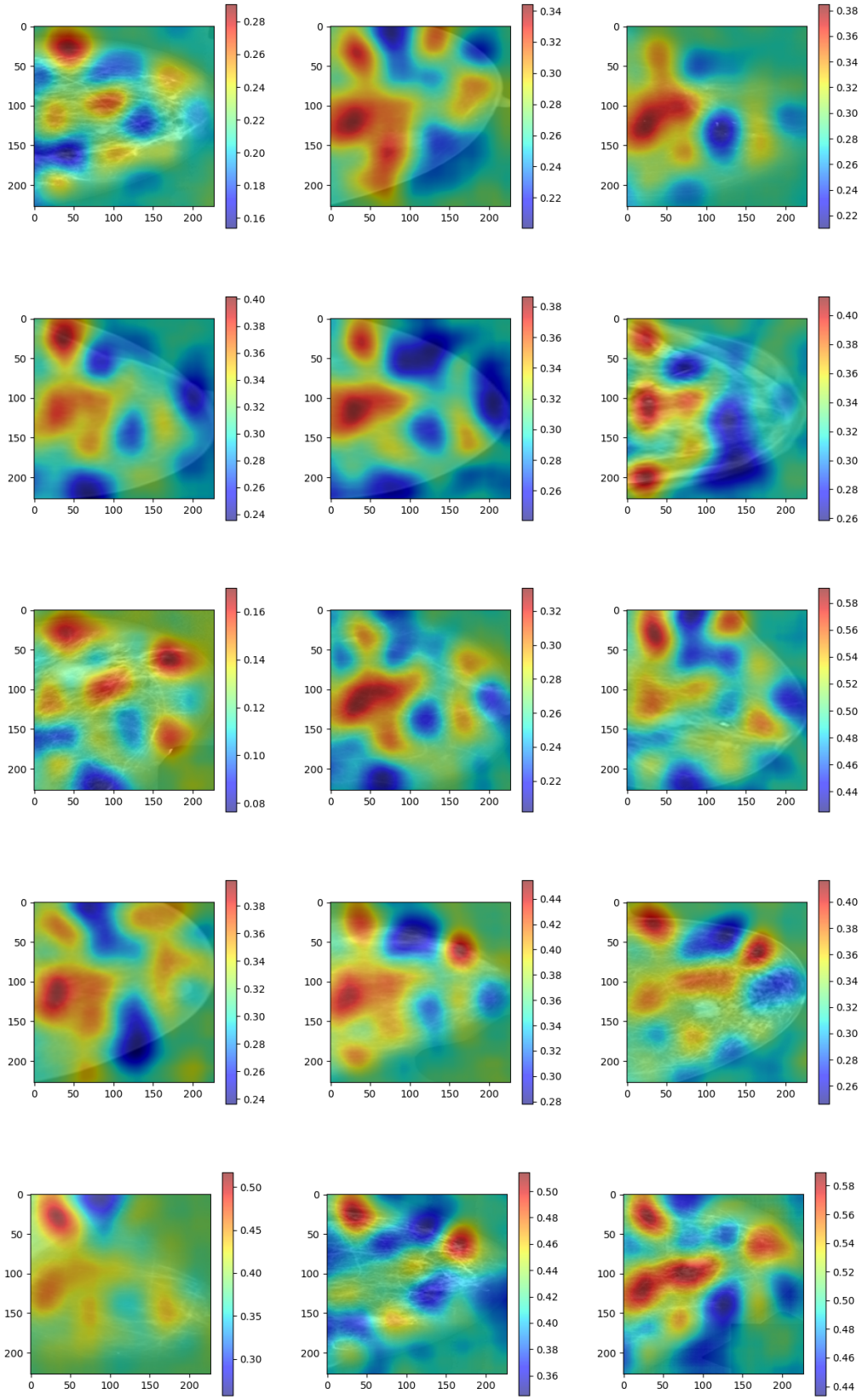


Figure B.2: Further Malignant RISE Explanations

# Appendix C

# Further SHAP Examples

This section contains further examples of SHAP explanations generated for images in the test set.

Segments with the highest impact on the classification decision are shown in green, with the least influential segments shown as red.

It's important to note that the SHAP value scales across these images are not consistent.

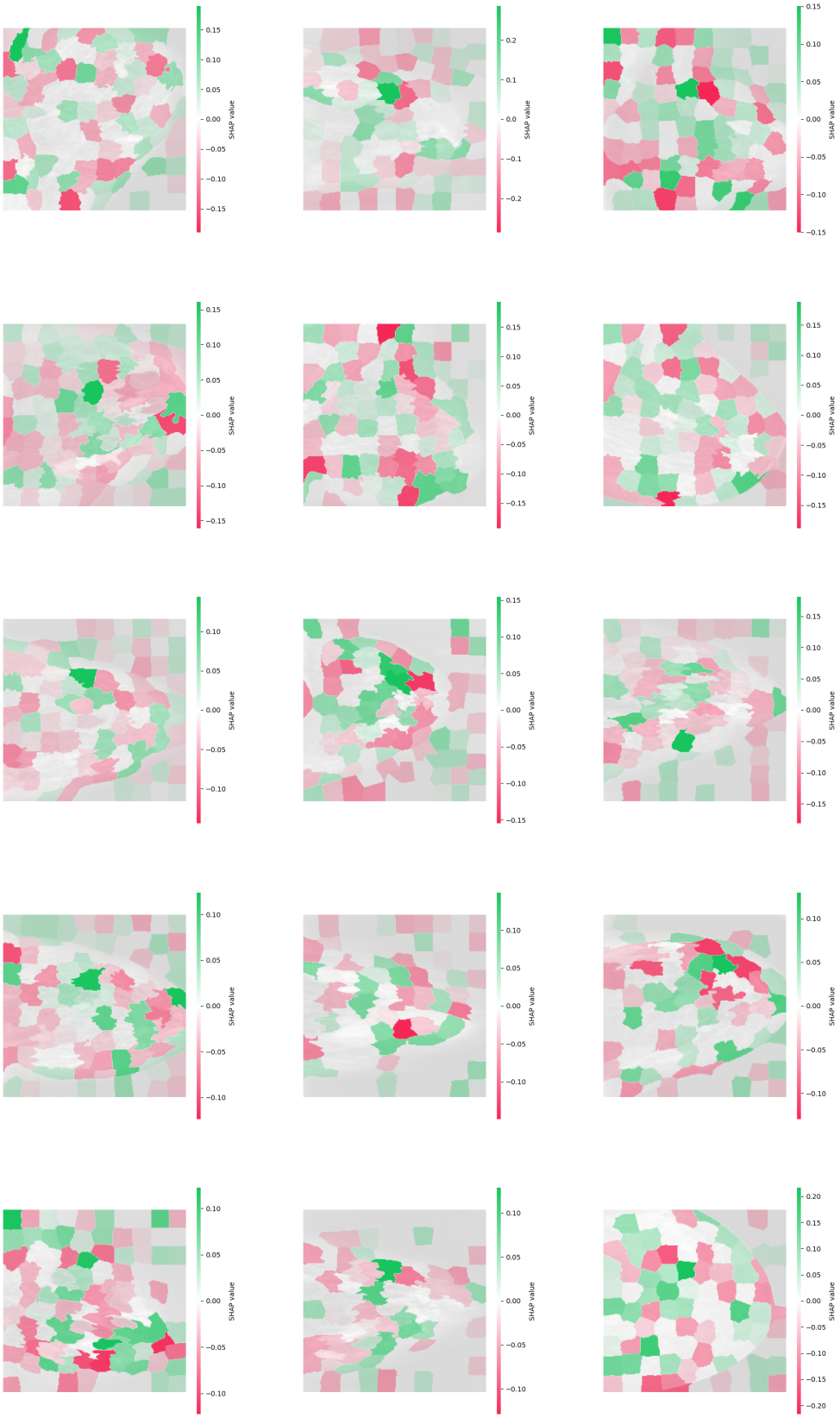



Figure C.1: Further Benign SHAP Explanations

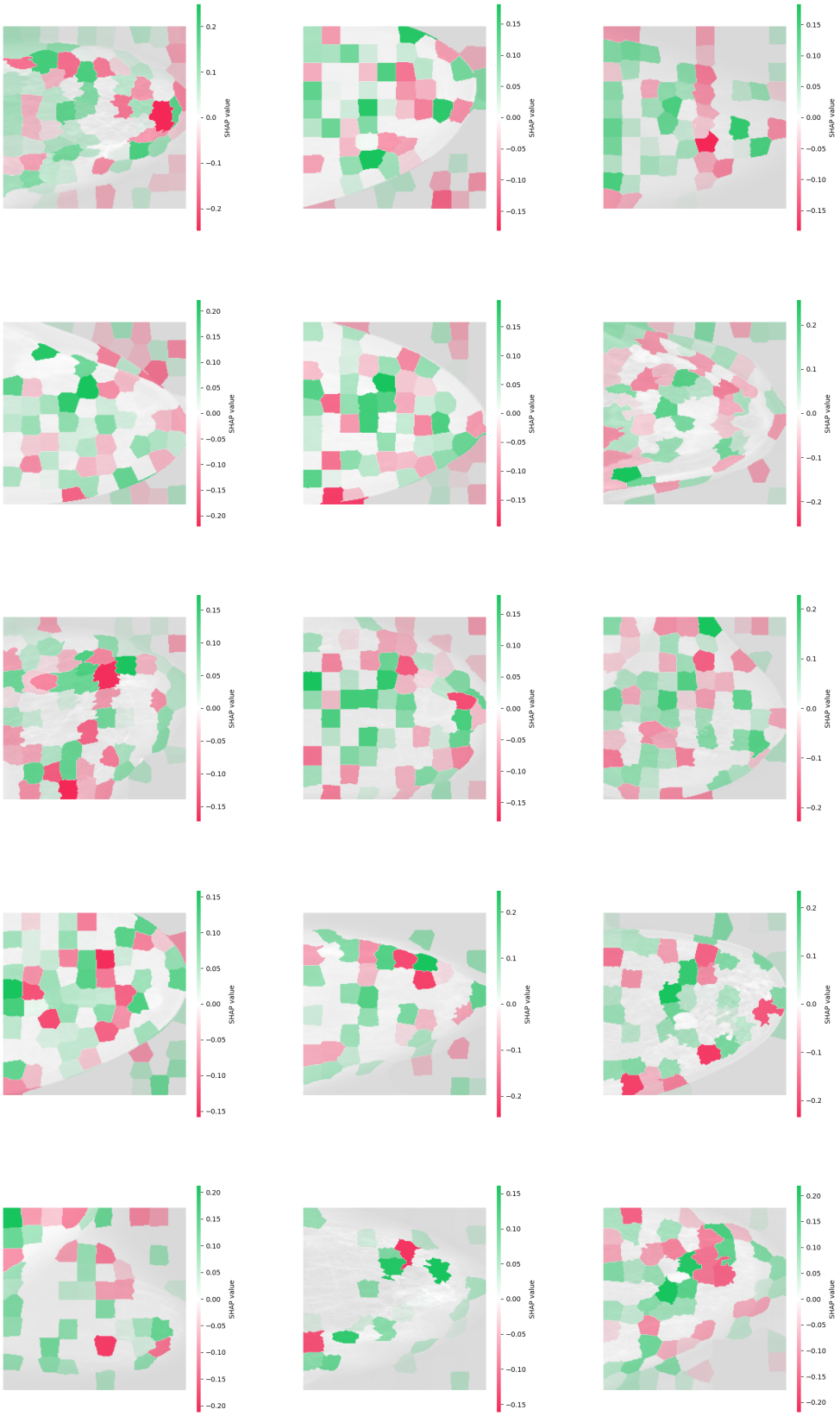



Figure C.2: Further Malignant SHAP Explanations

# Appendix D

# Further Pixel Comparison Visualisation Examples

This section contains further examples of the plots used in Chapter 5 to visualise the overlap, or lack thereof, of agreement with regards to the n most important pixels between LIME, RISE and SHAP. The value n is the number of pixels within the top 6 LIME features for an image - the process of empirically deciding this value is discussed in Chapter 4.

Results vary across images, however for all examples there are at least some areas of overlap.

The top n most important pixels as decided by SHAP are in blue, with RISE in red and LIME in green.

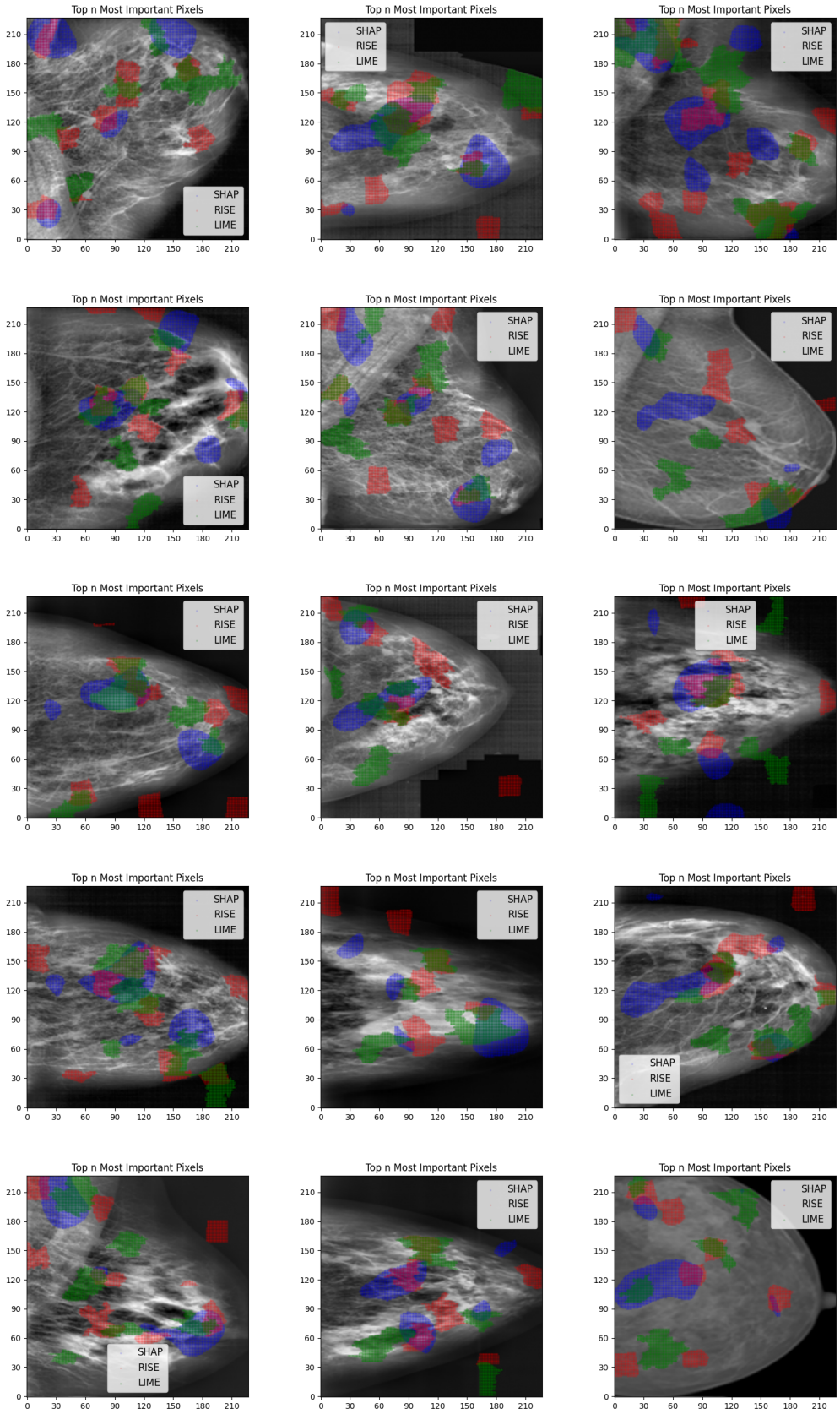



Figure D.1: Further overlay plots for Benign images

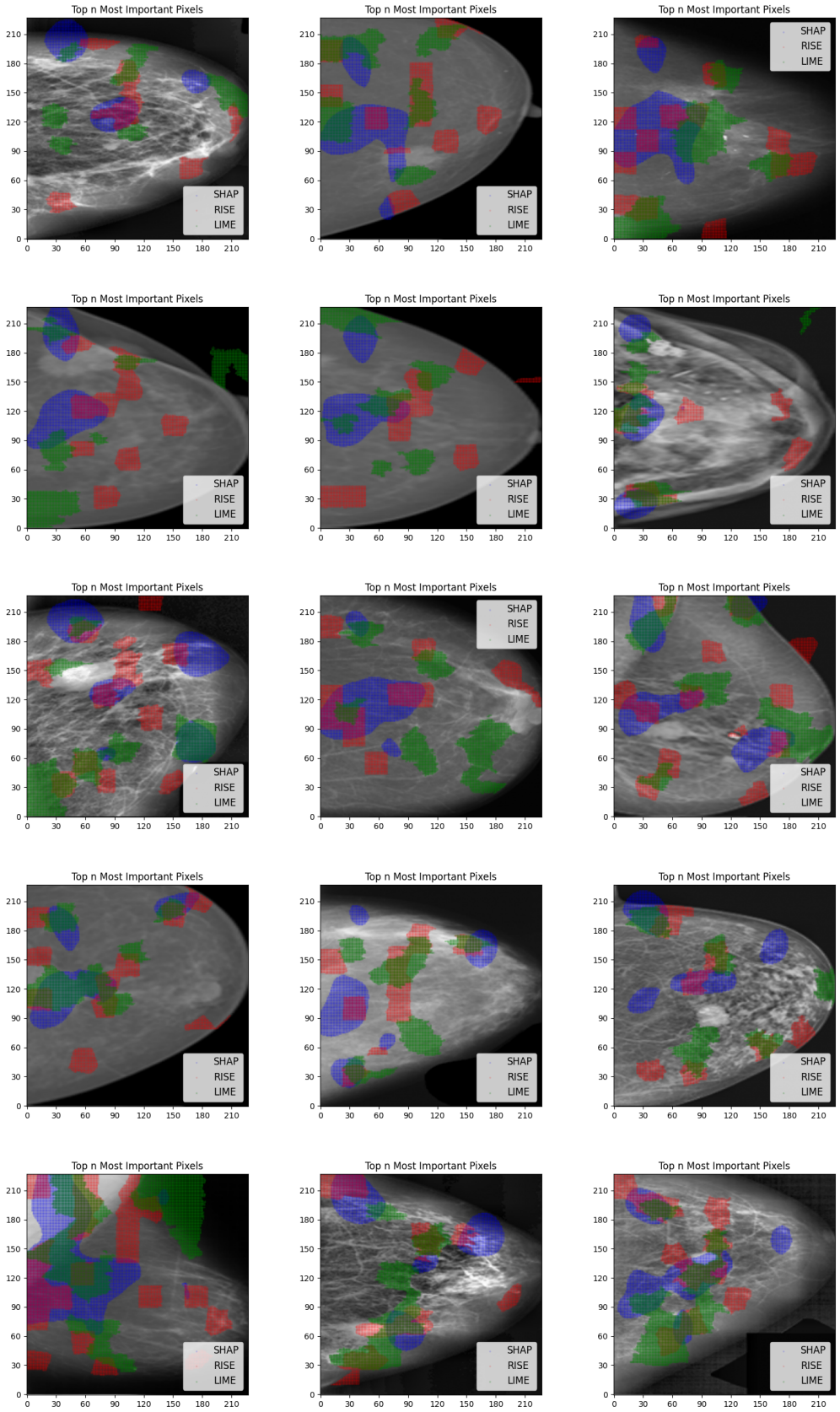



Figure D.2: Further overlay plots for Malignant images