Interpretability Approaches for a Breast Lesion Detection Model

Umaymah Imran

CMU-CS-22-137
August 2022

Computer Science Department
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Adam Perer, Chair
Ken Holstein

Submitted in partial fulfillment of the requirements for the degree of Master of Science.

© Umaymah Imran, 2022

Keywords: breast lesion detection, interpretability, radiomic features, deep learning, feature analysis

Abstract

Medical imaging holds an important role due to its ability to non-invasively visualize and analyze the internal structures of the human body. The rise in medical imaging data is putting an increased pressure on physicians/radiologists to efficiently perform clinical imaging tasks, which has in turn driven the need for development of diagnosis tools in healthcare. Therefore, over the last decade, there have been significant breakthroughs in the field of artificial intelligence for healthcare. Of these breakthroughs, the most important would be the development and deployment of deep neural networks (DNNs). These DNNs have a complex structure and consist of several computation layers. Their complex structure is what owes their ability to resolve challenging imaging tasks. However, several issues have been raised considering the black box nature of deep learning algorithms, which is why regardless of their impressive performance, DNNs have not achieved significant deployment in medical practice.

Specifically in healthcare, deep learning algorithms cannot be used for patient care unless the reasoning behind their outputs is explained due to the high stakes involved. Considering this, model interpretability is very important. It helps identify hidden information from the medical imaging data that may otherwise be invisible to the human eye. Model interpretability also adds to the trust of healthcare providers and patients involved. Beyond interpretability, analyzing model performance on targeted features could also allow model developers to debug their models.

In this thesis, we perform an analysis of a breast lesion detection model using different interpretability approaches. More specifically, we perform a preliminary analysis using the existing data and metadata. Then, we analyze hidden features for medical imaging data—that is, radiomic features—to further investigate model performance on certain input features. Finally, we determine the impact of the input data's location from the mammogram and suggest some future methods to improve our existing approaches.

Acknowledgements

I would like to thank my advisor, Professor Adam Perer, for guidance on the research direction, providing me with the relevant resources such as the UPMC mammogram dataset, giving useful feedback and helping resolve any problems I encountered throughout the process. Thank you for all your time and meetings, for helping brainstorm ideas with me, for looping in others for suggestions and for being super flexible with your availability.

I would also like to thank Ángel Alexander Cabrera for providing access to Zeno and related documentation, and for providing support in setting it up. Thank you for helping answer my questions, for always being available and for all your suggestions and ideas.

I would like to thank Bryan Lu for his assistance with understanding the model architecture and with accessing relevant resources such as the trained deep learning models, test data and related scripts, when needed. Thank you for clarifying many different aspects and for willing to meet whenever requested.

Contents

Introduction 9

Related Work 12

Feature Analysis 15
- 3.1 Dataset 15
- 3.2 Trained Models 17
- 3.3 Preliminary Analysis 17
- 3.4 Radiomic Analysis of ROIs 22

Investigating ROI Positions 37
- 4.1 Vertical Slices of Interest 41
- 4.2 Horizontal Slices of Interest 43
- 4.3 Analysis w.r.t. Horizontal and Vertical Position of ROI 45

Conclusion 46

Bibliography 48

Chapter 1

Introduction

In modern medicine, medical imaging holds an important role due to its ability to non-invasively visualize and analyze the internal structures of the human body. These medical images help identify and give more information about specific characteristics in the patient's targeted body area, which can be used to help with diagnosis for diseases, treatment planning, treatment follow-up and risk assessment. Due to advances in hardware technologies, increasing population and increasing awareness of medical imaging modalities, the amount of medical data is increasing rapidly. This means that there is an increased pressure on radiologists, clinicians and physicians in terms of managing to interpret and analyze this large amount of incoming data efficiently. Given the wide range of imaging data from different modalities and the fact that different people are working on their analysis and interpretation, variations have also been observed in performance of clinical imaging tasks [23]. Considering the large amounts of medical imaging data and the variations in analysis or interpretations of clinical imaging tasks, there has been a rising need in development of diagnosis tools in healthcare.

Over the last decade, there have been significant breakthroughs in the field of artificial intelligence for healthcare. This has multiplied the ability of computers to better comprehend and depict the given input. Of these breakthroughs, the most important one would be deep learning, which is based on development and deployment of deep neural networks (DNNs) [20]. These neural networks have a complex and deep structure, consisting of several computation layers. Their complex structure is what owes their ability to resolve challenging imaging tasks [18]. There are several different types of DNNs, such as Multi-Layer Perceptrons, Recurrent Neural Networks, Feed Forward Neural Networks and Convolutional Neural Networks, where each DNN has is more widely used for a specific use case. For the purpose of this thesis, we will be focusing on medical imaging—specifically, digital mammogram images.

Convolutional Neural Networks (CNNs) are the pillars of deep learning applications in medical imaging, which is what we also use for our application. Specific to medical imaging, CNNs have been recently employed more and have achieved state of the art results in image classification, image segmentation and localization, and disease detection and diagnosis [19].

As discussed, CNNs have become very popular for performing analysis of medical images over the past few years. However, several issues have been raised considering the black box nature of deep learning algorithms, which is why regardless of their impressive performance, DNNs have not achieved significant deployment in medical practice [19, 21, 22, 23]. Although DNNs are inspired by the human brain neurons, how these model work and derive their decisions is very different. These models can identify underlying correlations and features that are invisible to the human eye. More specifically, given the complex structure of CNNs, such as the several convolutional layers and nonlinear activations, it is difficult for humans to interpret and simply explain how the model produces a specific output from a given input [22]. Although theoretically it is possible, practically, it is not efficient.

Given the complexity of DNNs, it is also possible that the model learns false premise(s) on which its decision is based [18]. Even though this might not be a concern for some applications, it could be a serious issue in medical imaging because a wrong decision by the model could be detrimental to patient care. This includes, but is not limited to, late assessment of disease risk and incorrect diagnosis and/or treatment. Therefore, model interpretability is particularly important in medical imaging given the intricacy and high stakes of medical decisions [18, 19]. Results from the model should be explainable to a certain extent so that they are deemed reliable for integration with a Computer Aided Diagnosis (CAD) system for a physician's use and for assistance in their decision-making [18, 19]. This is because explanation for decisions is a crucial aspect to safe, ethical, fair, and trustable use and deployment of DNNs in real applications [21]. Without explanation of outputs, the limitations, biases and reasoning are hidden, which limits the utility of the DNN in the real world [23].

Medical diagnosis systems need to be transparent, understandable, and explainable in how they make their decisions for physicians, regulators and patients to have their trust in them [21]. Several new regulations, such as the European General Data Protection Regulation (GDPR) 2018, impose the requirement of explaining decisions by deep learning models in all sectors including healthcare [23]. Laws like these make model interpretability even more relevant since deep learning algorithms, as per such regulations, cannot be utilized for patient care unless the reasoning for their outputs is explained. More specifically, users should be able to understand the relationship between the features extracted by the model and its output for the DNN to be considered interpretable [22].

Considering the importance of model interpretability, tools that help interpret model output and help provide a basis for its decision need to be developed. In fact, several model interpretation techniques and tools have been developed by researchers to help better explain or visualize decisions by CNNs [19]. This allows to reduce the risk of making decisions on non-representative image features [18]. Interpretability of DNNs also helps identify faulty processes in prediction and can be used to troubleshoot prediction systems. It can help identify innate and hidden information from the medical imaging data that may otherwise be invisible to the human eye. Interpretability of DNNs also adds to the trust of healthcare providers and patients involved [23]. Moreover, beyond interpretability, such an approach could also allow model developers to debug models which may have learned incorrect image features. Through analysis, developers could determine what changes or fixes are needed to possibly address the issues with the model. For instance, this could involve addition of more diverse types of data in the model for lower performing subsets or slices of data. Therefore, in this thesis, we present an analysis of our breast lesion detection model, a deep convolutional model that is similar to VGG16. This model takes regions of interest (ROIs) or patches from digital mammogram images as input and outputs whether the patch has a lesion or not.

In this thesis, we make the following contributions:

- Understand where the model fails to correctly classify a patch by performing a preliminary analysis using the existing data and metadata
- Identify what hidden features, specifically radiomic features, in ROIs contribute to their classification and determine patterns for model performance based on feature distribution
- Determine ROIs from the mammograms on which the model performs best to identify regions for future patch extraction

Chapter 2

Related Work

There are several existing approaches for model interpretation, specific to mammogram classification. Model interpretation techniques can be broadly categorized under two main categories, that is, we can aim to understand: the model structure and functional layers or the model output predictions. Approaches that aim to understand the model structure and functional layers generally look at the hidden layers of the model [19]. These techniques explore the filters and features of the hidden layers, and they analyze the hidden representations of the data within the model [19]. In contrast, techniques that look at model output produce heatmaps or visualizations to understand the relationships between the different features that are important for the output [19]. For this thesis, we will be focusing on the second approach, that is, analyzing model output. This section describes some examples of studies that perform model interpretation using the model output on breast mammogram classification.

For breast mammogram classification, there are mainly two types of methods used for model output interpretation: attribution-based methods and non-attribution methods [19, 21].

Attribution-based methods measure how an input feature contributes to the target neuron, that is, to the output neuron for a classification task. Once we have this calculated for all features, then we can present these attributions in the shape of the input sample to get the heatmaps or attribution maps [21]. The advantage of using attribution-based methods is that they can easily be used on a black box, such as a CNN. This is because these methods are model agnostic, that is, they take a fully trained model as input, without having the need to understand or modify the underlying structure of the model [19, 21]. Moreover, there are several open-source implementations for these methods. This allows the deep learning developers to focus on designing an optimal model for a task and then using these off-the-shelf methods to understand the model better [19]. Due to these reasons, attribution-based methods are not only convenient, but are also powerful.

Saliency maps are an example of attribution-based approaches. Lévy and Jain [24] used two different CNNs, AlexNet and GoogleNet, to classify pre-segmented breast masses in mammograms as benign or malignant. For model interpretability, they visualized saliency maps to highlight areas in an image that drive the prediction of the CNN [22]. To do this, they computed the image's gradient with respect to the unnormalized class scores. Using the gradient value, the level of contribution was determined for prediction. For instance, regions with higher gradient values have higher contribution to prediction. They determined that both the CNNs were able to learn the edges of the mass and were sensitive to context, both of which are important aspects for diagnosis [24].

Another example of attribution-based methods is guided backpropagation [19]. Guided backpropagation is an extension of saliency maps. The main difference between the two is based on how backpropagation is conducted by the model through the ReLU activation layers, where ReLU is a commonly used activation function in CNNs. It applies ReLU to both the gradient of a neuron with ReLU activation and to gradient computation [21]. Shen et al. [28] uses guided backpropagation to visualize which areas of the input image were relevant for the predicted result by the classifier.

Non-attribution methods, unlike attribution-based methods, do not make use of existing implementations. Instead, these methods involve development of a methodology, which is then validated on a specific problem [21]. Examples of non-attribution approaches include attention maps, concept vectors, text justifications, generative modeling and combination with other machine learning methods etc. [21]. Although most of these techniques are model agnostic, similar to attribution-based methods, it is important to note that their implementation usually involves making modifications to the model architecture. For instance, for expert-based rule methods, we need to add expert knowledge. Since most methods in this category require development of a task-specific methodology and changes at the model's structural level, this adds a degree of complexity to implementation of these approaches. However, it is important to note that due to these tailored changes, we can potentially get more specific reasonings with the trade-off of higher effort [21].

Textual justification is a non-attribution approach. These models are designed to give textual justifications to the user to explain why they made a certain prediction given a specific input. Lee et al. [21, 25] developed a textual justification diagnostic LSTM model for breast mass classification on mammography images. It relied on input features to the classifier and the embeddings of predictions to give explanations and heatmaps for breast mass classification. The researchers used a visual word constraint loss to reduce the issue of duplication of sentences from the training data [23].

Another non-attribution approach are concept activation vectors (TCAV). Concept activation vectors give explanations as high-level image concepts for DNNs [21, 23]. More simply, testing with the TCAV method gives the users explanations for the different features learnt by the different layers in human understandable concepts [21]. The TVAC method quantifies the influence of a specific high-level image concept on the model's prediction. An extension on TCAV, called regression concept vectors (RCV), was made in medical imaging so that continuous variables, like radiomics (explained later), could be used as image concepts. In their study, Yeche et al. [26] used RCVs and a new metric called Unit Ball Surface Sampling (UBS). This metric is independent of the current layer and the representation of a concept in the feature space [21, 23]. Therefore, it could explain high-dimensional feature concepts across the different layers of the CNN used. Through use of RCVs, Yeche et al. determined what features were important for classifying calcification and masses from patches that had been extracted from mammograms. Thus, concept activation vectors can allow to explain high dimensional concepts across several layers, which can then be validated through input images.

The approach that we use for this thesis is similar to attribution-based methods, although with some exceptions. Similar to attribution-based methods, we employ a model-agnostic technique, that is, we conduct our analysis on a fully trained breast lesion detection model. Another similarity between our approach and the attribution-based methods is that we analyze the role of different features, that have been extracted from both the metadata and data, and how they contribute to a specific output by the model. However, different from heatmaps in attribution-based methods, our focus will be to analyze the distributions of the image features and identify any patterns common to incorrect or correct model predictions.

Chapter 3

Feature Analysis

3.1 Dataset

For our experiment, we use the mammogram dataset provided by the University of Pittsburgh Medical Center (UPMC), which has been deidentified. This dataset is comprised of 79501 digital mammogram images, that had been collected from approximately 22267 unique patients. As such, each patient has around 3 to 4 DICOM images, where each image is from one of the four standard views of the breast taken during mammography. The different views are L-CC (left craniocaudal), L-MLO (left mediolateral oblique), R-CC (right craniocaudal), and R-MLO (right mediolateral oblique), where the craniocaudal view is taken from top-down while mediolateral oblique is taken at a side angle. Figure 1 summarizes the distribution for the standard views of the breast taken during mammography.

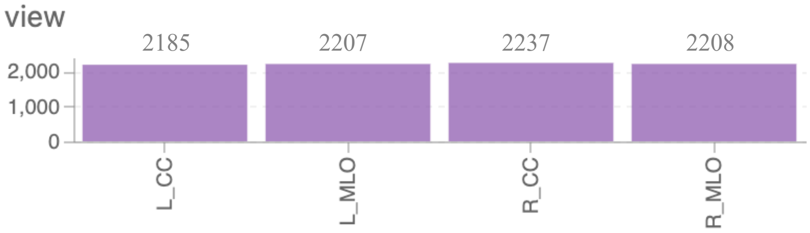


Figure 1 Distribution of Mammogram Views

The dataset comprises of Breast Imaging Reporting and Data System (BIRADS) scores 0, 1 and 2. Figure 2 summarizes the distribution for the BIRADS scores. BIRADS scores 0 and 2 correspond to mammograms that had been annotated by the radiologists, using ellipses to mark the region, due to presence of some abnormality or lesion(s). More specifically, BIRADS score 0 consists of mammograms with malignant findings, while BIRADS score 2 comprises of mammograms with non-malignant or benign findings. BIRADS score 1 corresponds to images that had normal breast tissue and since there was no suspicious breast tissue, these mammogram images had no annotations by radiologists.

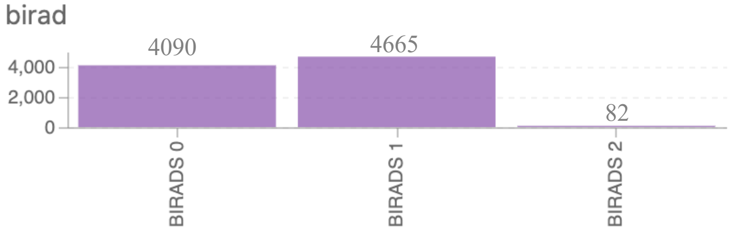


Figure 2 Distribution of BIRADS Scores

Given the data, a patch-based lesion classifier was developed. This CNN takes a patch from a mammogram image as input and determines if the patch contains a lesion or not. The model has an architecture similar to VGG-16. To generate the patches for this model, the no lesion or normal patches were extracted from BIRADS score 1 mammogram images, while the lesion patches were extracted from BIRADS scores 0 and 2. These were square patches of size 512 x 512, having a step size of 256. For BIRADS score 1, since there were no annotations by the radiologists, the region of interest was marked as a square region and was extracted from all tissue areas in the mammogram images. For BIRADS score 0, there were low-resolution annotated images by radiologists as well. These annotated images consisted of ellipses that encircled the region of interest. Therefore, the patches for these mammograms were extracted from the region within the ellipse. To do this, computer vision techniques were used to map the marked region in the low-resolution annotated images with the corresponding region in the high-resolution unannotated images to extract the patches. As for BIRADS score 2, a lot of the benign regions had not been annotated by the radiologists, which is why a significant portion of the mammograms from BIRADS score 2 were not used.

For the purpose of this experiment, we will be analyzing the performance of the model on the different patches having true labels as lesion (BIRADS scores 0 and 2) or normal (BIRADS score 1). We performed the analysis on 8837 patches in total, where 4090 patches were from BIRADS score 0, 4665 patches were from BIRADS score 1 and 82 patches were from BIRADS score 2. This is summarized by figure 3 below. We will be using the AUC score as our metric for the overall model performance and accuracy to analyze performance on a specific label.

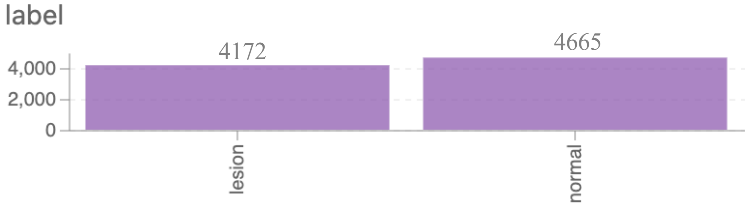


Figure 3 Distribution of True Patch Labels

3.2 Trained Models

For our analysis, we use a Convolutional Neural Network that has an architecture similar to a VGG-16. It is a patch-based model, which determines whether a patch extracted from a digital mammogram has a lesion or not.

We use two trained models: a best performing model and a normalized model. The best performing model is one which achieved the highest training accuracy (and AUC score). On the other hand, the normalized model is the one that was trained with z-normalization, so that the input has the same mean and standard deviation. Input patches for the normalized model have 0.5 mean and 0.5 standard deviation.

3.3 Preliminary Analysis

To understand how the model was performing on the patches, that is, what patches it was correctly classifying and what patches it was misclassifying, we used a framework called “Zeno”. For our preliminary analysis, we used the metadata and the data itself.

Currently under development by Carnegie Mellon University’s Data Interaction Group, Zeno[1] is a framework that aims to help debug and evaluate deep learning models. It allows the user to investigate and to analyze the model performance for their specific use case by feeding in the metadata, data and the trained models. The user can define specific preprocessing/distill or transformation functions that they might want to explore in terms of model performance. Zeno also allows its users to create “slices” using its interface. These slices or segments of data or information can help visualize targeted input, that is, input that meets certain conditions on which we want to filter. This in turn can help the user in understanding if there are any interesting patterns in the input data that might help determine cases where the model usually underperforms.

3.3.1 Features extracted from metadata

The metadata that we had available was very limited. Particularly, it consisted of the following features: the patch label, patient’s BIRADS score, the mammogram view from the four standard views (L-CC, L-MLO, R-CC, R-MLO), the patient’s age, and the starting and ending coordinates for the patches as extracted from the full digital mammogram.

For this purpose, we executed the analysis on our best performing model. Since the aim of the analysis was to determine the cases or the patterns where the model made wrong predictions, the focus of our analysis were the incorrect predictions. To do this, we defined 3 distill functions in Zeno to identify the incorrect predictions, the false positives and the false negatives. The false positive instances were cases where the original patch label was “normal” or “no lesion”, but the model incorrectly predicted that it was a “lesion” patch. The false negative instances were cases where the original patch label was “lesion”, but the model incorrectly predicted that it was a “normal” or “no lesion” patch. The incorrect instances comprised of both false positive and false

---

[1] http://zenoml.com

negative instances, that is, it consisted of misclassified instances by the model. Once we had these cases, the next step was to consider instances or ranges in the feature distributions, where the model was making most errors. Note that in comparison to previously analyzing the model's performance using accuracy only, we used the ROC analysis or AUC score as well. This is because many studies have suggested that AUC is a better metric in medical imaging, especially when there is an imbalance in data [14].

Considering the metadata features, there were a few simple comparisons that we performed. Generally, we found that the model performed slightly better on lesion patches in comparison to the normal patches. As for the BIRADS score, we found that the model performed best on BIRADS score 0, followed by BIRADS score 1 and finally, BIRADS score 2. However, the low performance on BIRADS score 2 was expected due to the limited annotated data available during model training. Therefore, the data for BIRADS score 2 was noisy in comparison to BIRADS scores 0 and 1. As for the DCM view types, the left views (L-CC and L-MLO) performed slightly better than the right views (R-CC and R-MLO). As for the patient's age, the model was performing better for patients over 60 years in comparison to patients 60 and below. This analysis has been summarized in the table 1. The analysis for the starting and ending coordinates for the patches is described in detail another section since the slicing is not straightforward, considering that the full digital mammograms have different sizes

Feature	Type	Model Performance *	No. of instances (%)
Patch Label	Lesion	88.49	4172 (47.2)
	Normal	86.37	4665 (52.8)
BIRADS Score	0	88.75	4090 (46.3)
	1	86.37	4665 (52.8)
	2	75.61	82 (0.9)
DCM View	L-CC	88.23	2185 (24.7)
	L-MLO	88.20	2207 (25.0)
	R-CC	87.49	2237 (25.3)
	R-MLO	85.83	2208 (25.0)
	L-CC + L-MLO	88.22	4392 (49.7)
	R-CC + R-MLO	86.65	4445 (50.3)
Age	<= 60	86.24	4668 (52.8)
	> 60	88.74	4169 (47.2)

* For patch labels, the model performance metric is accuracy. For all other labels, the metric is the AUC score.

Table 1 Metadata Feature-wise Model Performance

3.3.2 Features derived from data

Beyond analyzing the existing metadata features, there were some other features that were investigated as well. These features were not present in the existing metadata and so, were derived from the data itself for analysis. For instance, we wanted to explore how the model performed on the images considering their brightness and contrast. Therefore, we defined a distill function for each, where the function defined an image's brightness or contrast value.

Once we had preprocessed the images to determine their brightness, we analyzed the distribution for the image brightness values as in figure 4. We found that the mass of the brightness values for the images was close to the mean of the curve, that is, between the range 55 to 90 for brightness. However, for brightness values that were close to the tails of the distribution and away from the mean had an improved or worse performance—depending on the location. Images with high brightness, or with a brightness value of 90 or above, had a slightly decreased performance of 84.55% in comparison to the 87.37% overall performance. Images with high brightness consist of patches with high gray level values as shown in figure 5. These images have more visible "texture" than low brightness images. On the other hand, images with low brightness, or with a brightness value of 55 or below, had an improved performance of 90.27%. These images consist of patches with low gray level values as shown in figure 6 and "texture" is not as clear as it is in images with higher brightness.

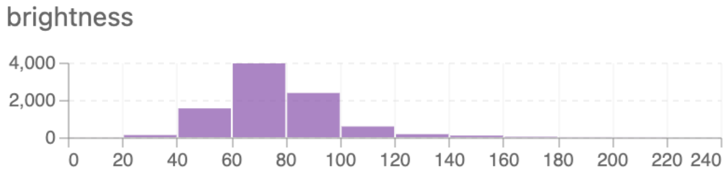


Figure 4 Distribution for ROI Brightness

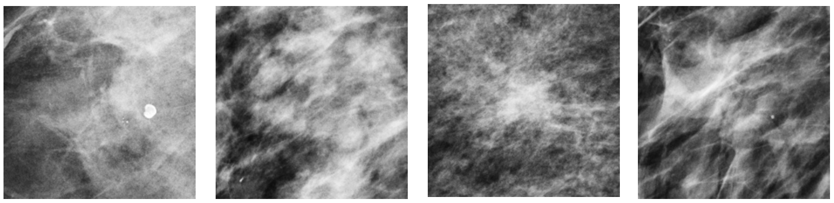


Figure 5 High Brightness ROIs

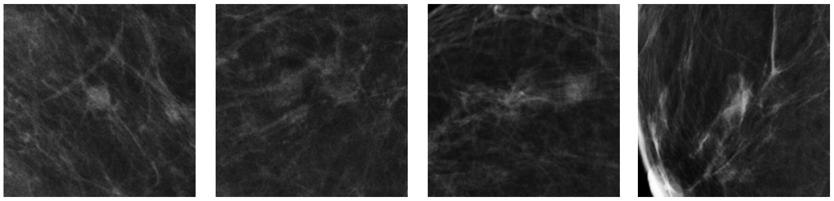


Figure 6 Low Brightness ROIs

Another feature that we derived from the data was the contrast level of the patch images. Contrast is difference in brightness between the light and dark parts of an image [17]. We found that the mass of the contrast values for the images was close to the mean of the curve, that is, between the range 18 to 42. However, contrast values close to the tails of the distribution and away from the mean had an improved or worse performance—depending on the location. Images with high contrast, or with a contrast value of 43 or above, had a decreased performance of 79.79% in comparison to the 87.37% overall performance. Images with high contrast consist of patches that have dense breast tissue as shown in figure 7. On the other hand, images with low contrast, or with a contrast value of 18 or below, had an improved performance of 93.46%. These images consist of patches with similar pixel values and an overall consistent appearance as shown in figure 8.

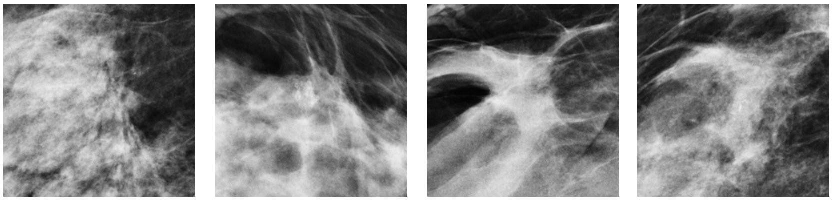


Figure 7 High Contrast ROIs

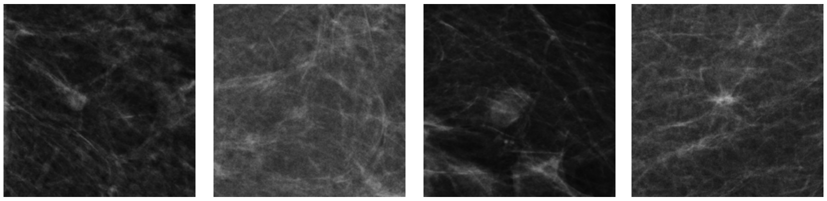


Figure 8 Low Contrast ROIs

While analyzing patches with low brightness, we observed that some patches had a plain black area. These images were patches that had been extracted from close to the bordering region of the breast and the black area was a segment from the background of the mammogram, outside the breast region. Therefore, we defined a distill function to observe these patches separately as well and determine any potential patterns in model performance. There were 1039 such patch instances and out of these, 665 patch instances had been extracted from normal mammograms. We observed that the overall performance of the model on patches with a black area was 90.09%, while for patches without a black area, the model's performance dropped to 87.01%. This change in model performance was mainly because of normal patches that had a black area, in which case the model had an increased accuracy. As discussed earlier, the overall accuracy of the model on normal patches is 86.37%, while it is 88.49% on lesion patches. We found that for normal or no lesion patches without a black area, the accuracy of the model decreased to 85.60% while for patches with a black area, the accuracy increased to 90.98%. For lesion patches with or without a black area, the accuracy of the model remained somewhat unchanged. We also compared the performance of our normalized model on normal patches with and without a black area. The overall accuracy of our normalized model on normal patches is 72.80%. Following a similar pattern to our best performing model, our normalized model performed significantly better on normal patches with black area(s) in comparison to normal patches without black area(s), having accuracies of 85.41% and 70.70% respectively.

We also performed different types of transformations on the patch images to see if it would have an impact on model performance. For this purpose, we analyzed both our best performing model and the normalized model, that is, the model trained with normalized patch images. One of the transformations we applied was a blurring filter on the patches using PIL. In case of blurring the images, the AUC scores of the models decreased. More specifically, both the models performed worse on lesion patches and better on normal patches. For instance, without the blurring filter, the best performing model had an accuracy of 88.49% on lesion patches and an accuracy of 86.37% on normal patches. After applying the blur filter, the accuracy of the lesion patches decreased to 84.95% and the accuracy of normal patches increased to 88.57%. Likewise, the normalized model had an accuracy of 93.31% on lesion patches and an accuracy of 72.80% on normal patches. After applying the blur filter, the accuracy of the lesion patches decreased to 90.92% and the accuracy of normal patches increased to 75.39%. Another transformation we applied was to see the effect of decreased brightness of patches. Although changing the brightness did not impact the accuracy of the normalized model, the accuracy of the best performing model decreased slightly on both lesion and normal patches, decreasing the model's AUC score marginally from 87.43 to 86.99.

Even though we extracted some features directly from the metadata, such as age and DCM view, and derived other features from the data, such as brightness and contrast, there were still patterns that were difficult or even impossible to quantify and represent directly from the medical images. Therefore, we used the PyRadiomics package to extract radiomic features from these patches.

3.4 Radiomic Analysis of ROIs

As mentioned earlier, using the given metadata and data, specific slices were created to better understand the underlying performance of the deep learning model on certain features. These features were either: (1) directly taken from the metadata, or (2) were derived from the data, that is, from the patches extracted from the mammograms. Features extracted from the metadata included the patient's age, BIRAD score or mammogram view type. Features derived from the data included an analysis of the brightness level and contrast of the image.

From these extracted features, there were seemingly interesting relationships between the predictions and the image characteristics. For instance, images with high brightness had a lower accuracy than images with a low brightness. Likewise, images with low contrast had higher accuracy than images with a high contrast.

Even though the extracted features revealed some interesting relationships between the predictions and the image characteristics, it is important to note that analysis of medical images is not that straightforward. This is because medical images contain a lot more information, such as radiological texture, that is not visible to the human eye [2]. As a result, in addition to extraction and analysis of metadata features and raw image features, several other hidden features or radiomic features were also extracted and analyzed as described in this section.

3.4.1 Relevant Work

Several researchers have proposed the need of performing quantitative studies on medical images to extract hidden information or image-based features [1, 2]. These quantitative studies on images have become possible due to several innovations in medical imaging resulting from advanced hardware, improved imaging agents, standardized protocols which allow quantitative imaging, and development of automated and reproducible imaging analysis methodologies [1]. More specifically, for this thesis, the most relevant component is the last component which focuses on innovations and improvements in image analysis; this is referred to as "radiomics".

As defined by Lambin et al. [1], radiomics is the high-throughput extraction and analysis of large amounts (over 200) of quantitative image features from radiographic images. It is an analytical framework that can be applied to different regions of interest or target locations and different imaging modalities such as MRI, US, DBT and mammography [2]. Radiomics assumes that the extracted, invisible imaging features are the result of interactions occurring at a more genetic and molecular level and are linked to the genotypic and phenotypic characteristics of the tissue [3]. Therefore, several studies have extended radiomics to correlate pathological or prognostic aspects with radiomic features because of the hypothesis that different tissue characteristics are represented as different radiomic values [2].

Unlike raw image features, radiomic features provide information on shape and spectral properties, tissue density, gray-scale patterns and inter-pixel relationships in radiological images [2]. Consequently, we can perform quantitative analysis of medical image data through radiomics to non-invasively identify intra-tumoral heterogeneity and to obtain more and better information than that given by a physician [1]. Specifically, in case of breast cancer or lesion detection, these quantitative or radiomic features can be used for lesion characterization [4]. This is because each patient is heterogeneous in their tissue texture, intensity or shape, which can be measured through radiomic featurization [1].

Several studies have conducted radiomic feature analysis on mammograms for several tasks, such as for lesion detection and for breast cancer risk assessment.

Mendel et al. [4] extract high-dimensional quantitative data or radiomic features from mammograms for temporal breast cancer risk assessment. Their study analyzes mammograms of patients who developed mammographic abnormalities over time, where the patients were categorized as having cancer or no cancer through biopsy. The researchers identified robust radiomic features from the mammograms. Then, they used these features to generate temporal features, which helped to identify the change(s) in the patient's breast tissue texture over time by comparing two mammographic time points of the patient. Finally, the changes in the temporal features were used to predict the patient's future risk of breast cancer.

Drukker et al. [5] investigate the combination of mammography radiomics, such as lesion texture, size, shape, and morphologic characteristics, and quantitative three-compartment breast (3CB) image analysis of dual-energy mammography to limit the number of unnecessary benign breast biopsies. Therefore, the main goal of their study was to decrease the number of false positive biopsy results, while ensuring that the true positive cases are not missed for biopsy. Overall, their approach of mammography radiomics combined with 3CB image analysis had a higher positive predictive value than conventional diagnostic digital mammography. However, there was slight loss in sensitivity.

Li et al. [6] explore whether combining radiomic features corresponding to normal contralateral breast tissue with radiomic features of breast tumors would improve the accuracy of lesion classification for diagnosis of breast cancer. They identified that their classifier with these combined radiomic features performed better than the one with lesion features only to differentiate malignant and benign lesions.

Karahaliou et al. [7] inspect whether texture properties of the tissue surrounding microcalcifications can contribute to breast cancer diagnosis. They concluded that texture analysis of surrounding tissue of microcalcifications, as in the patient's mammogram, has high potential to contribute to computer-aided diagnosis of breast cancer by limiting the number of benign unnecessary biopsies, while maintaining high sensitivity.

Tagliafico et al. [8] compare Digital Breast Tomosynthesis (DBT) of patients having normal breast tissue and dense breasts with patients having cancerous breast tissue. The researchers used radiomic features to evaluate normal and pathological breast parenchyma. Moreover, they investigated the correlations among the radiomic features and the clinical and prognostic parameters. They found that radiomic values for dense breast with normal breast tissue differed from cancerous breast tissue, which helps indicate that this new information could help in providing a better understanding of breast cancer detection through mammographic imaging and it could also contribute towards a more tailored screening and/or treatment for the patient.

Considering the wide range of work in radiomics specific to mammograms and the potential added information provided by the radiomic features, we decided to explore radiomic features for our use case as well. Our aim was to investigate correlations between the radiomic features calculated from the patches—that is, from the regions of interest—and the output by the model. Using this information, we were interested in determining what hidden information contributed towards classifying a patch as having a lesion or no lesion.

3.4.2 Steps to perform Radiomic Analysis

As discussed earlier, radiomic analysis involves analyzing high-dimensional features extracted from target sites in medical radiographic images. In this section, we describe the multiple steps involved in conducting radiomic analysis as shown in figure 9[2] [9].

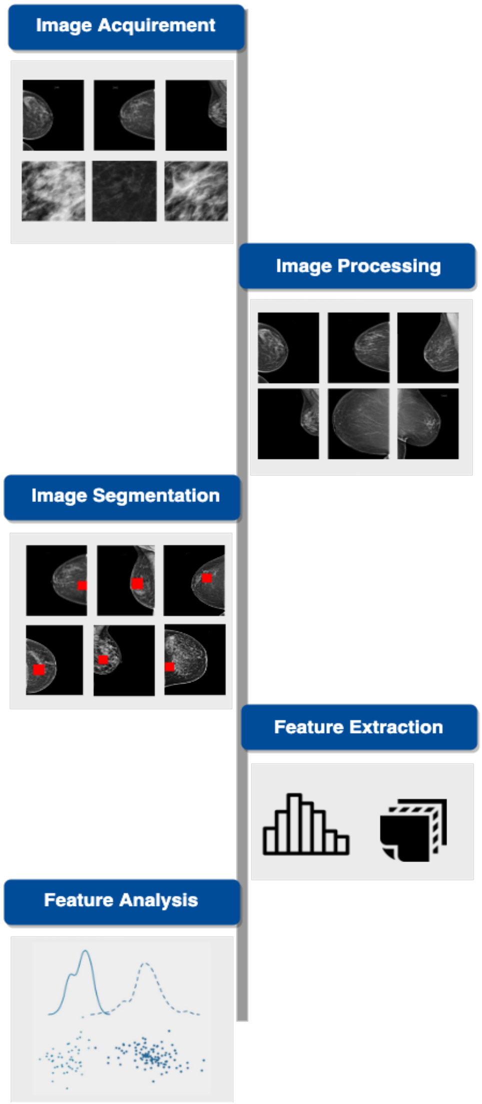


Figure 9 Radiomics Workflow

---

[2] https://www.flaticon.com/free-icons/texture
https://www.flaticon.com/free-icons/histogram

- **Image Acquirement**

The first step in conducting radiomic analysis was to acquire the appropriate images, that are high quality and standardized medical images [1].

In case of breast imaging, MRI, US, DBT and mammography are widely used, where each modality produces potentially different raw data due to different scanner models and imaging parameters. Due to these differences, the reproducibility of radiomic features is decreased and the difficulty in making direct comparisons among radiomic features is increased. Therefore, it is better to obtain all images from the same scanner and imaging parameters [2]. This experiment makes use of the UPITT mammogram dataset, which consists of 79501 breast images collected from approximately 22267 unique patients.

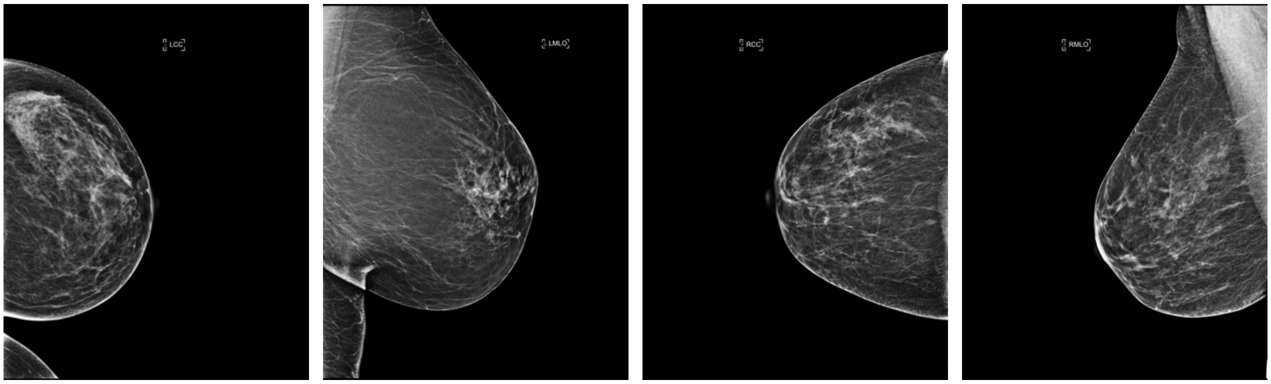
To perform an analysis of slices from basic metadata features and raw data (or patch) features, we initially used the patches in the test data itself. This is because the project uses a patch-based model for breast lesion detection and our initial approach was to determine if we could derive any interesting correlations between the patches and the predictions. However, this approach restricted the amount of information that was easily accessible—we could only rely of information derived from the metadata and the data itself. Therefore, we had to conduct radiomic analysis to access the hidden information from the medical images.

For radiomic analysis, we needed the full mammogram images from where the patches were extracted. Therefore, using the metadata of the patches in the test data, the corresponding mammogram images were obtained. These images comprised of both left and right breast images where each image was the high-resolution, unannotated mammogram. This was so that the labeling by the radiologists, as in BIRADS 0 and 2, did not interfere with the radiomic feature calculation.

- **Image Processing**

The next step in the radiomic analysis pipeline was image processing. This involved making certain transformations in the mammograms that were acquired from the first step.

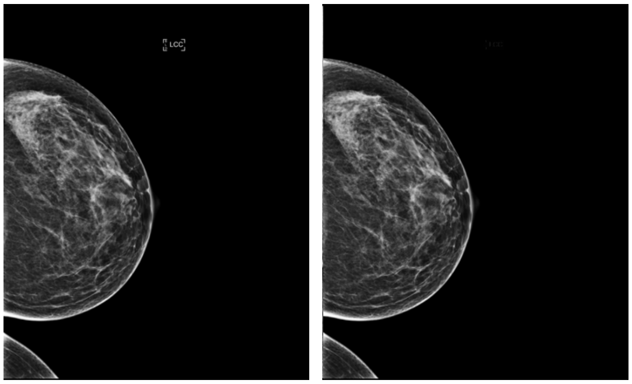
Figure 10 shows that mammograms had the view name, such as L-CC, L-MLO, R-CC or R-MLO, encompassed in a box and embedded within the image. Since radiomics determines the bounding boxes for the region of interest (ROI), this meant that these view labels or names could potentially interfere with the calculation of the radiomic feature values. To confirm this, we performed all steps in this section except for image processing to conduct radiomic analysis. We found that the view names did interfere with the feature calculations because it affected the coordinates for the bounding boxes of the ROIs. Therefore, these view names were removed from the images.



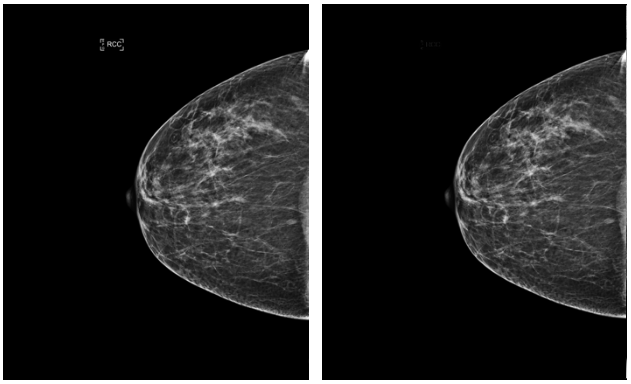
(a) L-CC (b) L-MLO (c) R-CC (d) R-MLO

Figure 10 Mammogram images with their DCM view labels

To remove the view names, the Python Imaging Library (PIL) was used to create copies of each mammogram image. Then, each image was modified such that the section of the image which had the view name was masked with black pixels. Since the mammograms contained the view of the breast, which had gray pixels, with a black background, this masking ensured that the labels of the view names merged completely with the black background. Figure 11 shows an example of both the left and right view of mammograms before and after the image processing step.



(a) L-CC (before) (a) L-CC (after)



(b) R-CC (before) (b) R-CC (after)

Figure 11 Examples of left (L-CC) and right (R-CC) DCM views before and after processing

- **Image Segmentation**

Once the images were acquired and processed, the next step in the radiomic feature pipeline was to perform image segmentation of our two-dimensional (2D) mammogram images. In image segmentation, we performed delineation of the region of interest (ROI) [9]. These ROIs define the target site from which the radiomic features are to be computed and limit the spatial extents of the radiomic analysis [2, 9]. ROIs can be defined manually, semi-automatically or automatically.

In case of mammograms with lesions from the UPITT mammogram dataset, we had mammogram images—with BIRADS score 0 and 2—that had ROIs annotated by radiologists, that is, with ellipses surrounding the regions with lesions. So, for our patch-based model, the patches were extracted from the corresponding non-annotated versions of the mammograms considering the location of the ellipse(s) using computer vision techniques. More simply, the patches for the lesions were image regions obtained from the full mammograms, that had been marked by a radiologist manually.

In case of mammograms from the UPITT mammogram dataset that had no lesions, we had non-annotated mammogram images for BIRADS 1. So, for our patch-based model, the patches were extracted from the breast tissue area in the normal mammograms.

Considering how patches from both lesion and no lesion mammograms were extracted, our ROI specification was performed semi-automatically, where annotations by radiologists was performed manually and patch generation was executed automatically. Many studies have demonstrated that the ROI size is an important factor in medical imaging for diagnosis [4, 12, 13, 14]. These ROIs were square sites of size 512 x 512 pixels. The starting x and y coordinates of these patches, as in the full mammogram images, were stored in the metadata as well.

To generate the radiomic features for these ROIs, we used the PyRadiomics[3] library, which is an open-source Python package for extracting radiomic features from medical imaging. This package allows for feature extraction from our 2D mammograms and calculates single values per radiomic feature for each ROI.

As required by the PyRadiomics package, in order to generate radiomic features for an instance, we need both the source image and the segmentation or masked image, where both the source image and the segmentation image should have equal dimensions. The source image should be the processed digital mammography image from the previous step, while the segmentation or masked image should be an image with the ROI masking on digital mammography image.

Since our segmentation images were patches of size 512 x 512 pixels, while our mammograms were images of sizes 3328 x 4096 pixels or 2560 x 3328 pixels, the next step in the image segmentation step was to transform the patches into usable images for PyRadiomics. To do this, there were several steps taken. First, for each processed digital mammography image, a copy was created. Next, considering the location of the patch—that is, its x and y coordinates—in the digital mammography image, a masked image was created with the ROI. To do this, a red region was marked on each mammography image from where the patch was extracted to mark the ROI. This

---

[3] https://github.com/AIM-Harvard/pyradiomics

red region was a label for the ROIs, allowing PyRadiomics to identify the masked area or ROIs in the segmented image. Figure 12 shows an example of a source digital mammography image and its corresponding segmentation image.

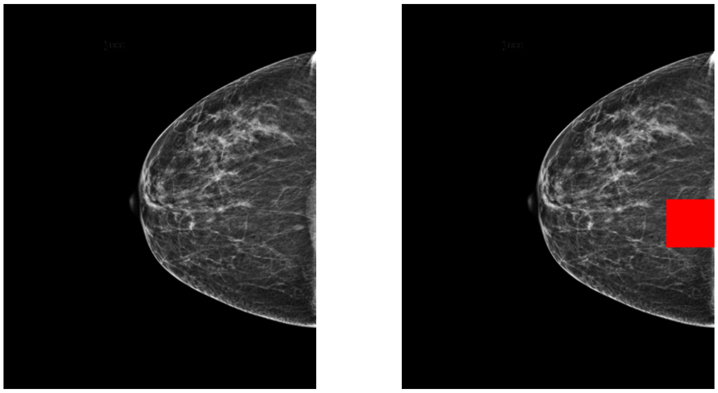

Original Mammogram Image
Masked Mammogram Image

Figure 12 Example of an original mammogram and its masked duplicate, with the ROI marked

- **Feature Extraction & Analysis**

The fourth step in the radiomic analysis pipeline was feature extraction and analysis. This involves computing the quantitative imaging features or radiomic feature from the ROIs, where each feature is defined by a mathematical formula [2].

Once the radiomic features were calculated, we used our normalized lesion detection model for analysis. This model was trained on normalized patch images. The overall accuracy of the normalized classifier is 82.48% (AUC 83.05), where the accuracy on the lesion ROIs is 93.66% while the accuracy on the normal ROIs is 72.24%. Although the features were extracted for all ROIs, note that the analysis was focused on ROIs extracted from normal digital mammograms since the model significantly underperforms on normal ROIs. Therefore, the radiomic analysis will be focused on ROIs extracted from normal mammograms.

Radiomic features can be grouped into four main categories, that are, morphological, histogram-based, textural, and transform-based features [2].

Morphological features relate to the shape and physical attributes of the ROI [2]. However, note that in our application, morphological features are not considered since our ROIs do not precisely mark or draw the boundary around a lesion or site. Instead, our ROIs are focused on a square area, where each segmentation image has the same ROI shape and dimensions. Therefore, morphological features are not applicable.

Histogram-based or first order statistics relate to information from the intensity histogram of the ROI and do not consider the spatial relationships [16]. There are several prominent features in histogram-based features, for instance energy, variance, kurtosis, skewness, median, interquartile range, entropy, uniformity. Several breast imaging studies have suggested that entropy, mean, minimum and maximum are important features [2]. For our analysis, we analyzed some of the important histogram features to identify the ranges or values for specific features from the ROIs where the model was underperforming as shown in table 2. We have described some of the most meaningful features in more detail below.

Entropy is the amount of randomness in pixel values of the image. As described in the documentation of PyRadiomics, it determines the average information needed to encode image values. In our case, when entropy is low, error rate is low and when entropy is high, error rate is high. More specifically, for lower entropy values, the AUC of the model rises from 83.05 to 85.99. So, for patches with entropy values between 0 and 2 as in figure 13, the accuracy of the model rises significantly for the normal patches, that is, from 72.24% to 78.78%. Beyond this given range, as for patches in figure 13, the model performs worse and the overall AUC score of the model drops from 83.05 to 79.73. Specifically, the accuracy for normal patches drops to 65.69% and drops further to 58% for entropy values beyond 2.6.

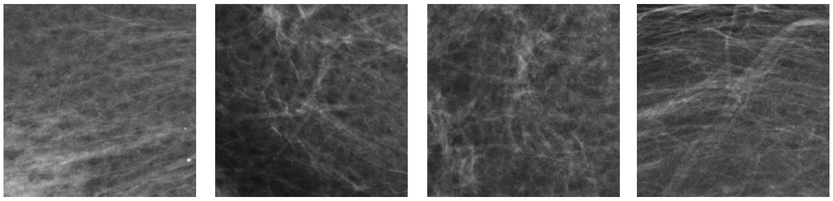


Figure 13 Normal patches with low entropy (between 0 and 2)

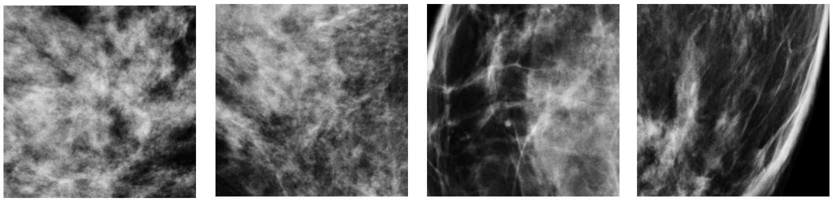


Figure 14 Normal patches with high entropy (> 2)

Mean is the average value of an ROI's gray level intensity. After an analysis of the mean distribution, we found that the model performs better on patches with low mean values. For mean values 60 and below, the AUC of the model rises to 88.21 from 83.05. This is because for normal ROIs with a low mean as in figure 15—that is, with a mean lower than 60—the model performs very well and has an accuracy of 81.25% in comparison to 72.24%. The highest number of normal patches are between the ranges of 60 and 80, where the accuracy remains around 71-72% for normal patches. However, for patches with a mean beyond 80, as in figure 16, the accuracy further drops to around 66.45%. Therefore, ROI images that have less variation in terms of gray level intensities tend to perform better than images with greater variations in the gray level intensities.

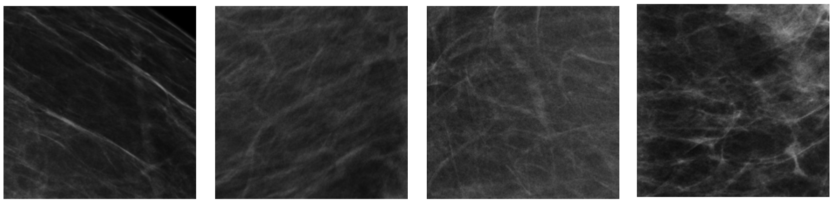


Figure 15 Normal patches with low mean (< 60)

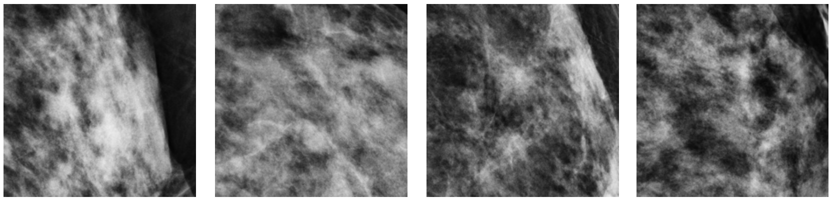


Figure 16 Normal patches with high mean (> 80)

Maximum is the highest value for the ROI's gray level intensity. For normal ROIs with a low maximum gray level intensity value as in figure 17, the accuracy of the model is high and reaches around 79% for intensities 200 and below. However, beyond a maximum value of 200, as in figure 18, the overall accuracy for normal patches decreases to 68.61%.

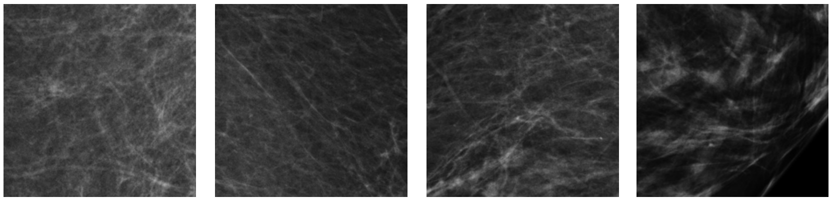


Figure 17 Normal patches with low maximum (<= 200)

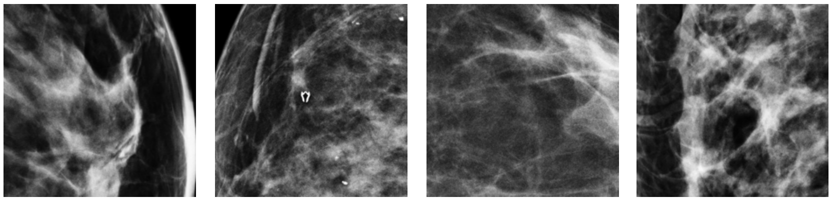


Figure 18 Normal patches with high maximum (> 200)

Minimum is the lowest value for the ROI's gray level intensity. For normal ROIs with a low minimum gray level intensity value as in figure 19, the accuracy of the model is low and reaches around 62.96% for intensities 15 and below. However, beyond a minimum pixel value of 15 as in figure 20, the overall accuracy for normal patches increases to around 74%.

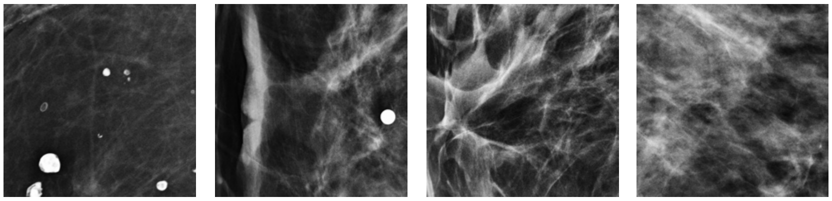


Figure 19 Normal patches with low minimum (<= 15)

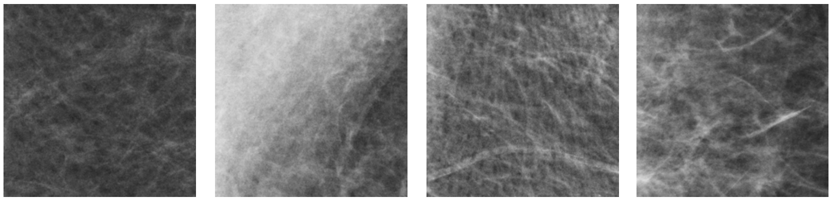


Figure 20 Normal patches with high minimum (> 15)

Considering the analysis for the minimum, maximum and range features, all of which consider the minimum and/or maximum gray level intensities in the ROI, we found that the model performed better for normal patches that had a smaller difference or range between the maximum and minimum intensity value. When range is high, then performance decreases.

Uniformity determines the degree of homogeneity in the image. This means that a high value for uniformity means higher homogeneity or in other words, a smaller range for distinct intensity values. After analysis for uniformity, results showed that the model performed worse for normal patches with low uniformity—especially below 0.3, where accuracy dropped to 61%. Less uniform images have more heterogenous regions, that is, what appears as texture. Normal patches that are more homogenous have a higher accuracy.

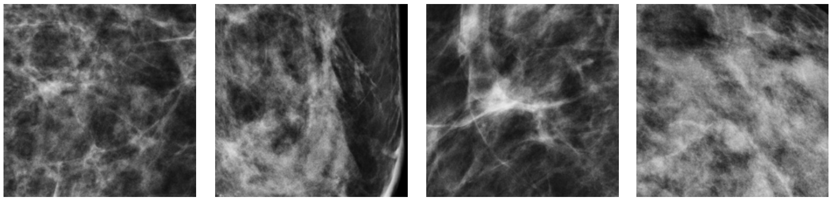


Figure 21 Normal patches with low uniformity (< 0.3)

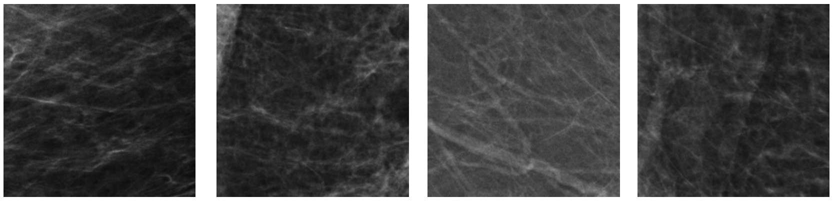


Figure 22 Normal patches with high uniformity (> 0.5)

Other than first order statistics features, we have second order statistics features as well. Second order statistics features are more commonly referred to as the textural features [2, 16]. Textural features consider the neighborhood information of the voxels. These are features that are most important and have driven radiomic studies [2]. These features can be derived from the gray level co-occurrence matrix (GLCM), the gray level run-length matrix (GLRLM), the gray level size zone matrix (GLSZM), the gray level dependence matrix (GLDM) and the neighboring gray tone difference matrix (NGTDM) [29]. For our experiment, we will be focusing on selected features from GLCM, as it is the most commonly used category for radiomics in breast mammography [6]. GLCM essentially represents intensity pairs and quantifies their frequency in the neighborhood [2, 16]. We will also be analyzing selected features from NGTDM, since it helps better analyze the difference between a pixel's gray intensity value and the average gray values of its neighbors.

An important feature from the GLCM is contrast. This feature measures the variations in intensities within the ROI. Analyzing the distribution for contrast reveals that as value for contrast gets larger, the model's accuracy for the normal patches decreases. For instance, after 0.4 as in figure 24, the accuracy on normal patches decreases to 69.87%. A larger value for contrast correlates with a greater difference in the intensity values of the neighborhood regions. Therefore, as the intensity differences between the neighborhood regions increase, the accuracy on a normal patch decreases.

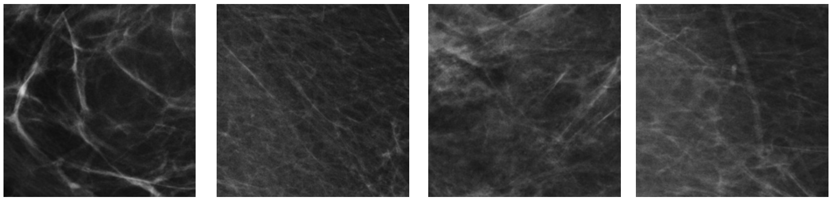


Figure 23 Normal patches with low contrast (> 0.25)

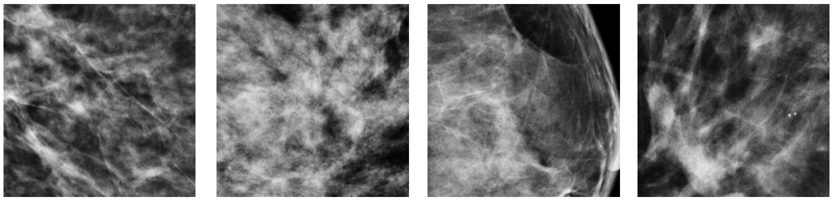


Figure 24 Normal patches with high contrast (> 0.4)

Correlation is another important feature in GLCM. It determines image linearity, representing linear dependency of gray level values to their respective voxels in the GLCM [29]. A value of 0 means uncorrelated, while a value of 1 means they are strongly correlated. When the value of correlation is high, the accuracy of the model on normal patches decreases. For instance, nearly half of the normal instances under analysis have a value of 0.85 and above for correlation. However, for a correlation value above 0.85, as in figure 25, the accuracy of the model on normal patches decreases to 67.29%, while the performance of the model on normal instances is around 77% for correlation values below 0.85 as in figure 26. Therefore, in case of normal patches, higher the correlation value, lower the performance.

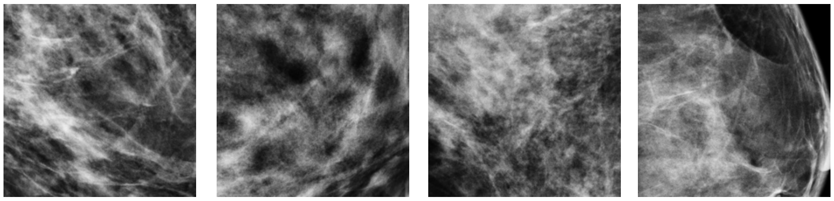


Figure 25 Normal patches with high correlation (> 0.85)

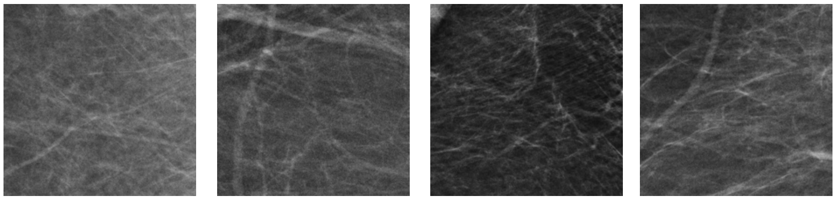


Figure 26 Normal patches with low correlation (< 0.85)

In GLCM, entropy is an important feature. There are several features that focus on entropy, these include joint entropy, difference entropy and sum entropy. Joint entropy measures randomness in the intensity values of the neighborhoods. Difference entropy measures the amount of randomness in the differences between the neighborhood intensity values. Sum entropy measures the amount of randomness in the sums of the neighborhood pixel intensities. In case of all three entropy-related features, as the randomness or variability in the neighborhood intensities increases, the model's performance on normal patches decreases. Normal patches with high entropies have an appearance similar to patches in figure 25, while normal patches with low entropy have an appearance similar to patches in figure 26.

Joint energy is a feature particularly relevant to image homogeneity. A higher value for joint energy implies that there are more homogenous regions in an image. We found that the model performs generally better on homogenous images. For joint energy values above 0.15, the model has an AUC score of 85.78, with an accuracy of 93.21% on lesion patches and 78.36% on normal patches. Figure 27 shows examples of normal patches with joint energy values above 0.15. However, for joint energy values 0.15 and below, the model's AUC score drops to 79.31. While the accuracy on lesion patches is unaffected, the performance of the model on normal patches drops to 64.71%. Figure 28 shows some examples of normal patches with joint energy values 0.15 and below. Therefore, the model does not perform well on heterogenous normal patches.

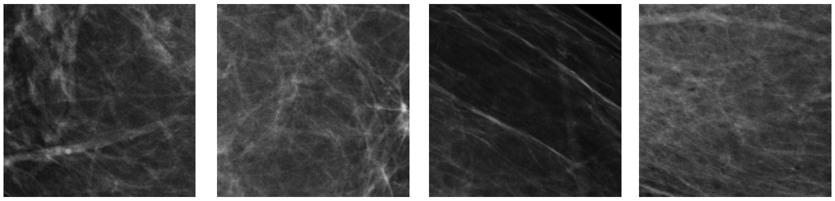


Figure 27 Normal patches with high joint energy (> 0.15)

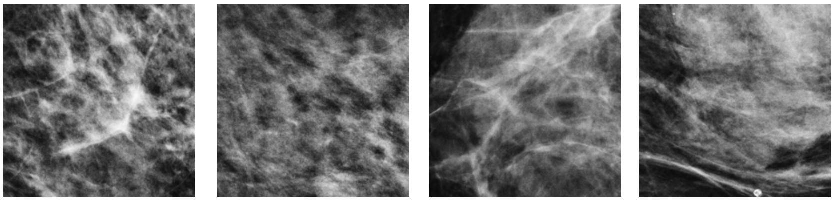


Figure 28 Normal patches with low joint energy (<= 0.15)

Information measure of correlation 1 (IMC1), information measure of correlation 2 (IMC2) and maximum correlation coefficient (MCC) are all features that quantify the complexity of the ROI texture [30]. As for IMC1, values -0.40 and below give a poor performance on normal patches as the accuracy drops to 67.83% even though around half of the normal patch instances fall in this range. For IMC2, as the value increases, the accuracy of the model on normal patches drops. Although there are more than one-third of normal patch instances for IMC2 values above 0.90, the accuracy of the model is only 65.81%. A higher value for MCC means a more complex texture. Therefore, as the value of MCC increases, the accuracy on normal patches decreases. Particularly, there are 1390 normal instances with an MCC value between 0.85 and 0.95 and the accuracy of the model between this range is only 66.62%. Figures 29 and 30 are examples of normal patches for these features, where the model performs better on patches from figure 30.

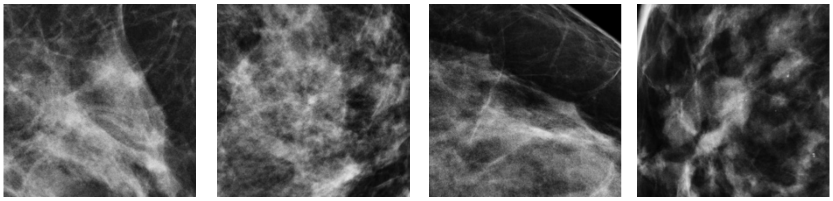


Figure 29 Normal patches with low IMC1 (< -0.40) + high IMC2 (> 0.90) + high MCC (> 0.85)

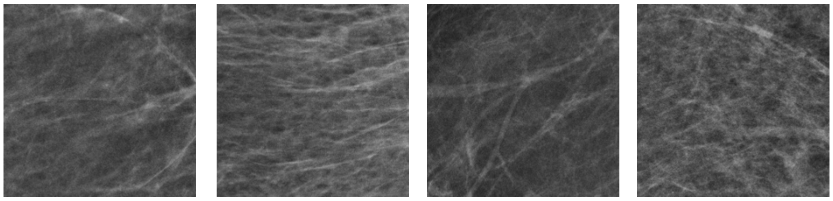


Figure 30 Normal patches with high IMC1 (> -0.30) + low IMC2 (< 0.80) + low MCC (< 0.75)

Another important feature in GLCM is sum average. Sum average measures the overall brightness of the image. Similar to our inference in the non-radiomic slice analysis, we found that the model performs better on images with a lower brightness than images with higher brightness. The brightness value of most normal patches was between 6 to 7. For normal patches with a sum average over 7, which were one-third of the total normal patches, the model accuracy dropped to 65.36%. However, for normal patches with a sum average below 6, which were one-third of the total normal patches, the model accuracy improved to 80.00%. Therefore, for the current model, the brighter the images, the worse the performance.

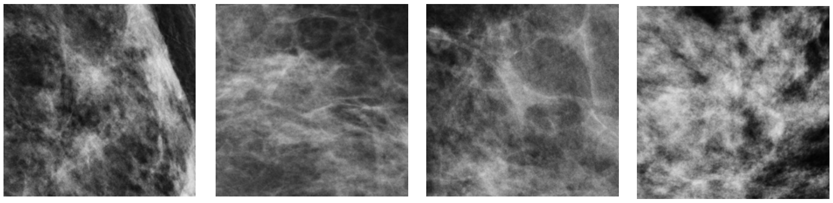


Figure 31 Normal patches with high sum average (> 7.00)

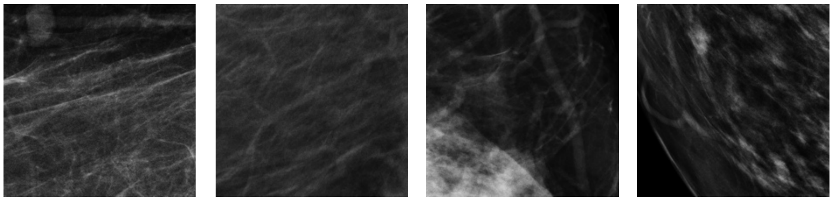


Figure 32 Normal patches with low sum average (< 6.00)

As for NGTDM, a relevant feature is coarseness for analysis. As the name suggests, coarseness refers to how coarse an image is, that is, it is a measure of the spatial rate of change [29]. As the value of coarseness increases, there is a wider range of gray level intensities in the ROI. Therefore, a high value of coarseness results in a lower model performance. For instance, the overall accuracy on normal patches with a high coarseness value, that is of 0.00006, or above is 63.67%. Figure 33 shows some examples of normal patches with high coarseness. In contrast, the model performs better on normal patches with low coarseness, having an accuracy around 78%. Figure 34 shows some examples of normal patches with low coarseness.

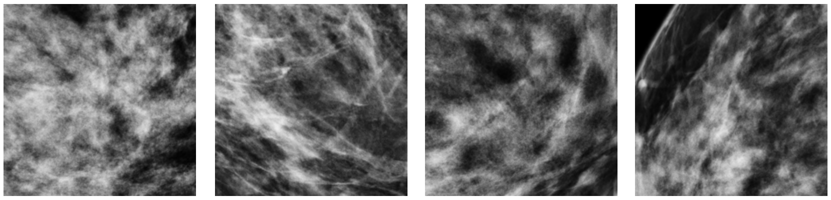


Figure 33 Normal patches with high coarseness (> 0.00006)

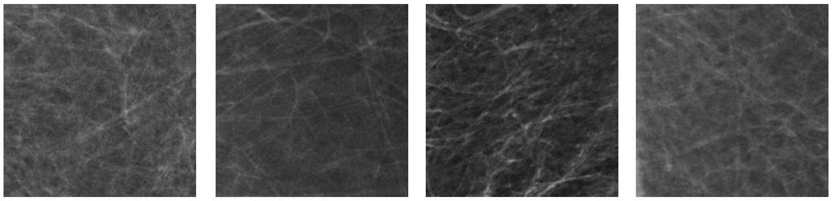


Figure 34 Normal patches with low coarseness (< 0.00005)

Complexity is another feature of interest from NGTDM. It is a measure of the number of primitives in an ROI image. So, a complex image is one which has a higher number of primitives, that is, the image is non-uniform and rapidly changing gray level intensities [29]. After analysis of ROIs, more complex ROIs are those that have more appearance of texture or dense breast tissue. As reflected by many other radiomic features, when a normal patch has rapidly changing gray level intensities, then the model tends to make more errors. In this case, for complexity values over 12, that is for patches as in figure 35, the model accuracy on normal patches is around 68% only. On the other hand, for lower complexity values, as for patches in figure 36, the model accuracy on normal patches is around 79%.

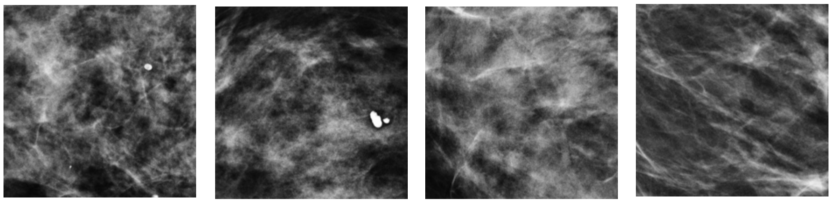


Figure 35 Normal patches with high complexity (> 12.00)

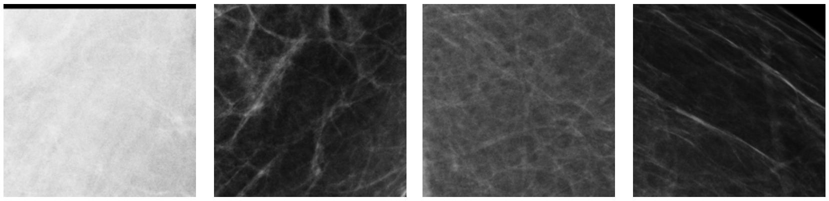


Figure 36 Normal patches with low complexity (< 10.00)

The contrast feature in NGTDM considers the changes in the neighborhood intensities and the overall gray level range of the image. Therefore, contrast is high when there are large changes in the neighborhood intensities and range of gray level intensities is also high. Normal patches that have a high contrast value tend to have poorer model performance. More specifically, nearly half of the normal patch images have contrast values over 0.015. For these images as in figure 37, the overall model accuracy is 65.46%.

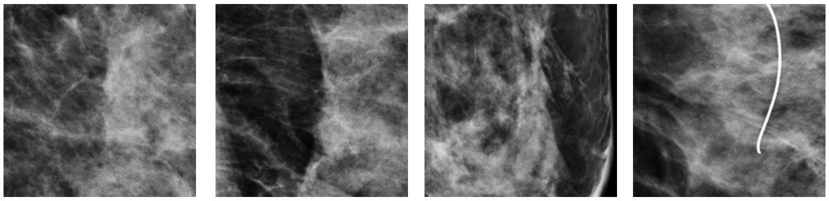


Figure 37 Normal patches with high contrast (> 0.015)

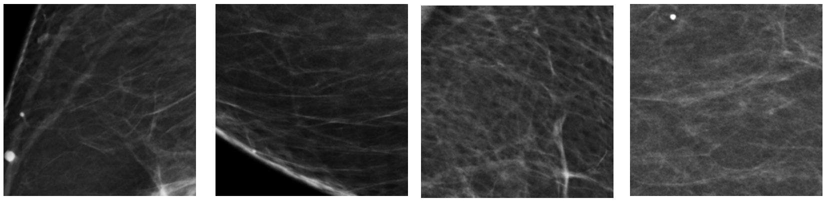


Figure 38 Normal patches with low contrast (< 0.0075)

Given the analysis for features in table 2, there were several radiomic features that helped better explain ROIs where the model was unperforming—particularly for the normal patch images. Unlike the slices analyzed in the previous section, radiomic analysis helped better see the relationship between the different neighborhoods in the image which further allowed to investigate the relationship between the model performance and texture better.

Feature Category	Type
Histogram-based (First Order)	Entropy Mean Maximum Minimum Uniformity
GLCM (Second Order)	Contrast Correlation Joint Entropy Difference Entropy Sum Entropy Joint Energy Information Measure of Correlation 1 (IMC1) Information Measure of Correlation 2 (IMC2) Maximum Correlation Coefficient (MCC) Sum Average
NGTDM (Second Order)	Coarseness Complexity Contrast

Table 2   Radiomic features analyzed

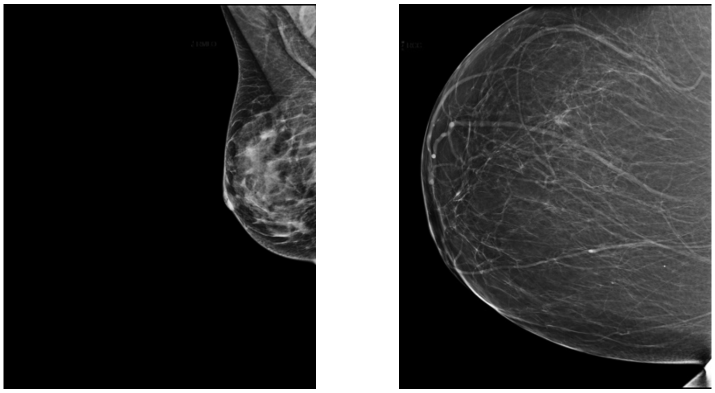
Chapter 4

Investigating ROI Positions

Many previous studies have shown that ROI placement plays an important role in medical imaging since it contains essential diagnostic information [12, 13, 14]. For instance, women at a higher risk of breast cancer usually have more dense breasts than women at a lower risk [14, 15]. Therefore, considering this, several studies have selected ROIs from the central breast region—that is, from the region immediately behind the nipple because this region tends to have the densest breast parenchyma. Focusing on this region has resulted in the highest performance for distinguishing between low-risk and high-risk women and in determining abnormalities or gene mutations in patients [4, 12]. These studies have also shown that varying the ROI location and extracting it outside of the central breast region results in a decreased performance of the deep learning model predictions or computation of texture features [14].

As discussed in the previous sections, our ROIs in mammograms with lesions were sites where the expert radiologists had made their annotations and our ROIs in mammograms without lesions were sites that were tissue regions in normal mammograms. Since the ROIs with the lesions were manually marked by the radiologists and the normal ROIs were arbitrarily taken, it was not necessarily the case that the ROI was a central breast region—it could be patch from the top, bottom or a corner region. Therefore, it is important to analyze how the current model is performing on the ROIs outside the central breast region, as suggested by other studies.

In order to perform this analysis, there were several factors that needed consideration. First, the mammograms consisted of both left and right view images. This meant that the relative position of the patch in the mammogram was dependent on the breast view. Second, the covered region by the breast in each mammogram differed from another. More specifically, while one breast could cover 20% of the digital mammogram image, another could cover 70% of the image as shown in figure 39 (a) and (b), respectively. This means that a coordinate region which might be a central breast region for some mammograms might be the top or bottom breast region for other mammograms. Third, the sizes of the mammograms were different, where some were of size 2560 x 3328 pixels and others were of size 3328 x 4096 pixels. Similar to the previous point, we could not define a strict coordinate region to correspond to a specific breast region.



(a) (b)

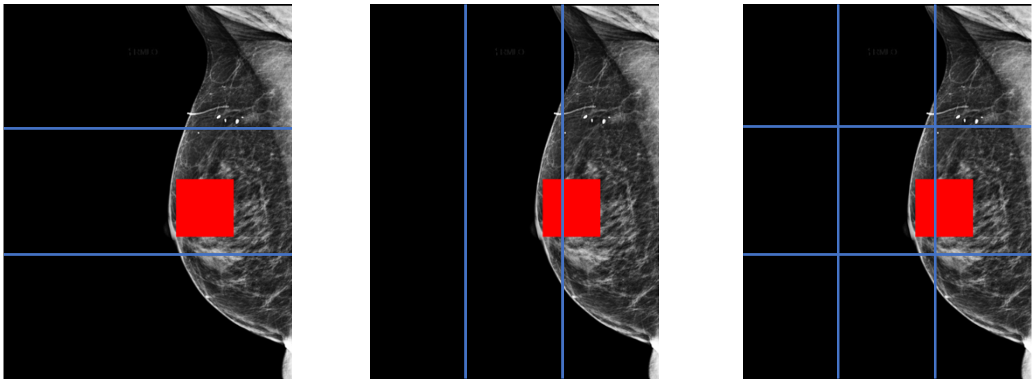
Figure 39 Mammograms have a significantly different percentages of image region covered by breast tissue

To tackle the above-mentioned issues, we wrote a script that sections the digital mammogram image into horizontal and vertical sections or slices based on the image dimensions. It also determines in which [horizontal, vertical] pairing the ROI is located in. So, given a fixed number of horizontal and vertical slices, we use the starting x and y positions of a patch and the original dimensions of its corresponding mammogram to determine the patch's position in the original mammogram—that is, the horizontal and vertical slices that the patch lies in.

To classify an ROI's location in the horizontal and vertical slices grid, there are two possible cases: (1) an ROI could lie entirely in the region of a horizontal or vertical slice, or (2) it could lie (or overlap) between two horizontal or vertical slices.

In the first scenario, the ROI would be categorized as part of the horizontal or vertical slice that it completely lies in. For instance, as in figure 40 (a), since the patch lies entirely in horizontal group 1, the patch's horizontal position is 1.

In the second scenario, the ROI's horizontal or vertical position depends on the ROI's area in each of the overlapping slices. The ROI is categorized as part of the horizontal or vertical slice where it has a higher covered area. For instance, as in figure 40 (b), the patch is present in both vertical groups 1 and 2. However, since the patch has a higher covered area in vertical group 2, it will be categorized as part of vertical slice 2.



(a) Horizontal Slices (b) Vertical Slices (c) Grid Slices

Figure 40 (a) – (b) shows how a mammogram is sectioned into 3 horizontal slices and 3 vertical slices. In this case, given the horizontal and vertical slices, the mammogram can be further represented as 9 smaller slices or visually, as a 3 x 3 grid (c), where each [i, j] in the grid corresponds to horizontal slice i and vertical slice j.

In order to explore what coordinate regions (in terms of horizontal and vertical slices) are of most interest, we defined all possible horizontal and vertical slices considering the ROI size 512 x 512 pixels and the mammogram dimensions 2560 x 3328 pixels and 3328 x 4096 pixels. Note that the minimum width of a horizontal or vertical slice has to be 512 pixels, therefore, the maximum number of horizontal or vertical slices is dependent on the result = image dimension divided by 512. Given that, the maximum number of horizontal slices is 6 and the maximum number of vertical slices is 5.

We defined the number of horizontal slices between 3 to 6 (both inclusive). Figure 41 shows the distribution of all horizontal slices. For instance, “3horizontal_region” in figure 41 means that the mammograms were divided into 3 horizontal coordinate regions and the patch was categorized under one of the following groups: 0, 1 or 2. Similarly, “4horizontal_region” in figure 41 means that the mammograms were divided into 4 horizontal coordinate regions and the patch was categorized under one of the following groups: 0, 1, 2 and 3.

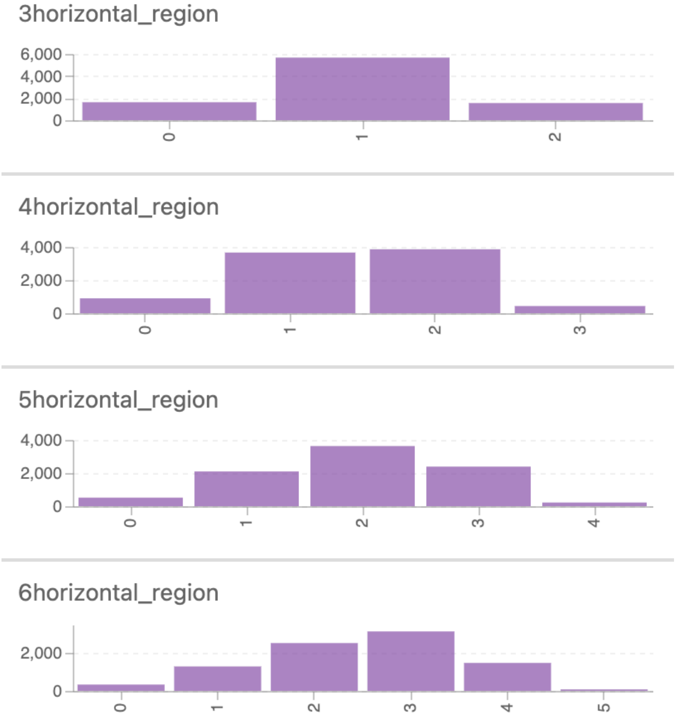


Figure 41 Distributions for the different horizontal slices

We defined the number of vertical slices between 3 to 5 (both inclusive). Figure 42 shows the distribution of all vertical slices. For instance, “3vertical_region” in figure 42 means that the mammograms were divided into 3 vertical coordinate regions and the patch was categorized under one of the following groups: 0, 1 or 2. Likewise, “4vertical_region” in figure 42 means that the mammograms were divided into 4 vertical coordinate regions and the patch was categorized under one of the following groups: 0, 1, 2 or 3.

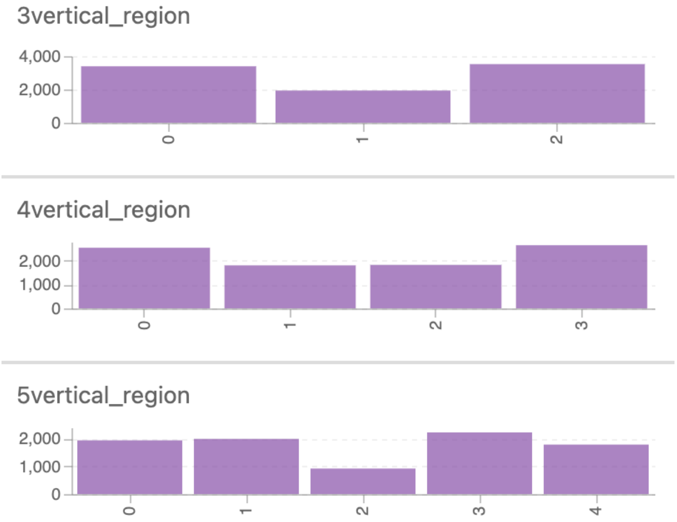


Figure 42 Distributions for the different vertical slices

To choose the slices of interest, we first filtered out all ROIs that had a plain black region present in them. This feature was described under the metadata features in chapter 3. The purpose of filtering such ROIs was to eliminate any patches that had been extracted from close to the outer boundary of the breast in the full mammogram because any boundary region does not correspond to the central breast region. After this filtration, we had a total number of 7798 patches remaining. The next step following this was to determine which horizontal and vertical slices were optimal for being categorized as the central breast region. For this purpose, we used our best performing model for analysis.

4.1 Vertical Slices of Interest

After analyzing the distributions for vertical slices 3 to 5, we found that the ideal slicing for the vertical slices was 5 sections. This means that to analyze central breast patches, it is best to vertically segment a mammogram into 5 vertical regions as in figure 43 (a), where the number marking each slice is the index position or category number for the vertical slice.

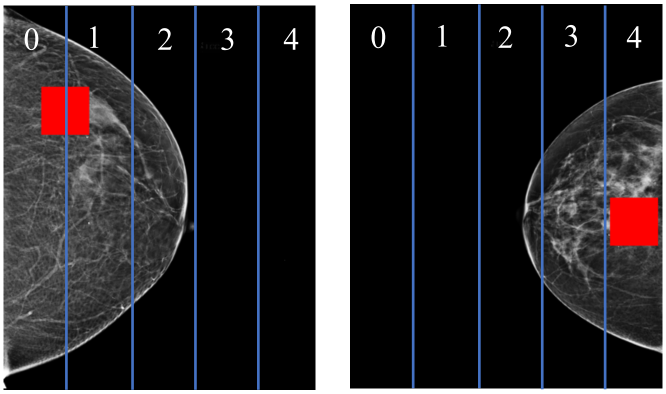


Figure 43 (a) Vertical slicing of mammograms into 5 segments

For 3 vertical slices, most of the ROIs from the left mammogram views were concentrated under category 0, while most of the ROIs for right mammogram views were concentrated under category 2. ROIs under category 1 comprised of the frontal outer breast tip region for both the left and right mammogram views. Given 3 vertical slices, it was difficult to determine if the ROI had been extracted from the central region of the left or right view. This slicing more or less just returned whether the ROI was from a left view or a right view. Therefore, we further sliced the mammograms into 4 vertical slices.

For 4 vertical slices, most of the ROIs for left mammogram views were grouped under vertical categories 0 and 1, while most of the ROIs for right mammogram views were grouped under categories 2 and 3. This is because for left views, most of the breast region corresponds to vertical index positions 0 and 1; the same logic follows for right views. Since the ROIs were almost entirely divided into only one of 2 categories for each view, it was difficult to distinguish which ROIs came from the central breast region. It was only possible to determine if the ROI was more towards the left or right of the breast region in the mammogram images.

Considering the limitations of 3 and 4 vertical slices, we divided the mammograms into 5 vertical slices. This slicing performed best in comparison to the previous two vertical slices in terms of categorizing the breast regions. ROIs extracted from the left mammogram views were concentrated between categories 0 to 2, while ROIs extracted from right mammogram views concentrated between categories 2 to 4.

As discussed earlier, the mammograms comprised of both left and right breast views. Therefore, the central vertical breast regions would be a union of the corresponding central vertical slice for the left views and the corresponding central vertical slice for the right views. To determine which vertical slice best corresponds to each one of left or right view, the AUC scores for the most relevant slice category was compared with the rest of the categories. Then, the slice category that outperformed the other categories was selected for that mammogram view.

For the right views without the plain black region, we have an AUC score of 86.27 and 3904 instances. We selected vertical category 3 from the 5 vertical slices to correspond to the central breast region for right views. In comparison to all other categories, category 3 had the highest number of instances/ROIs and the highest AUC score of 87.15. Regions outside of category 3 had a combined AUC score of 85.34. These results are summarized in table 3.

Vertical Category	AUC of category	Combined AUC of other categories	No. of Instances
0	79.71	86.39	69
1	85.04	86.31	87
2	83.42	86.55	371
3	87.15	85.34	1797
4	85.78	86.44	1580

Table 3 Right Views: Analysis of AUC scores for vertical slices

For the left views without the plain black region, we have an AUC score of 87.79 and 3894 instances. We selected vertical category 1 from the 5 vertical slices as the central breast region for left views. In comparison to all categories, category 1 had the highest number of instances or ROIs and the highest AUC score of 88.68. Regions outside of category 1 had a combined AUC score of 86.69. These results are summarized in table 4.

Vertical Category	AUC of category	Combined AUC of other categories	No. of Instances
0	86.80	88.27	1647
1	88.68	86.69	1712
2	84.31	87.98	370
3	87.30	87.86	118
4	100	87.72	4

Table 4 Left Views: Analysis of AUC scores for vertical slices

As reflected by the AUC scores for both left and right views, it is evident that ROIs extracted from outside the vertical central breast region perform worse in comparison to the ROIs extracted from the vertical central breast regions—categories 1 (left views) and 3 (right views).

4.2 Horizontal Slices of Interest

For both the left and right views, the horizontal slice would be a single region where the central breast region would be located. After analyzing the distributions for horizontal slices 3 to 6, we found that the ideal slicing for the horizontal slices was 5 sections. This means that to analyze central breast patches, it is best to horizontally segment a mammogram into 5 horizontal regions as in figure 43 (b), where the number marking each slice is the index position or category number for the horizontal slice.

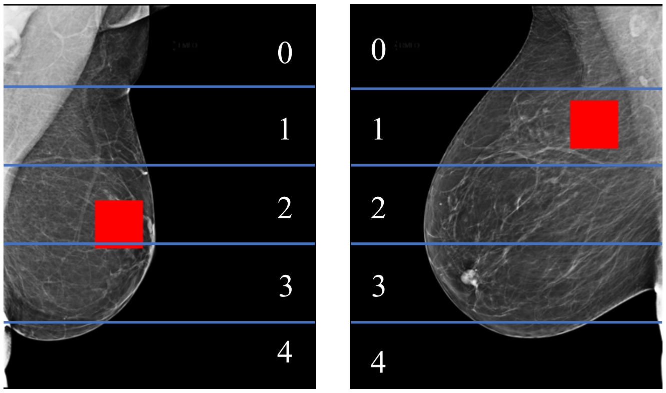


Figure 43 (b) Horizontal slicing of mammograms into 5 segments

To get the central breast region in a mammogram, we first sectioned the mammograms into 3 horizontal slices. For most mammograms, from categories 0, 1 and 2, category 1 corresponds to the central breast region. This horizontal region had an AUC score of 85.60, with 5313 total number of ROIs. However, we wanted to narrow this region further to target the region behind the nipple of the breast. Therefore, we sliced the mammograms further into 4 slices.

Instances in category 1 from 3 horizontal slices corresponded to instances in categories 1 and 2 in 4 horizontal slices as in figure 44, with category 1 having an AUC score of 85.35 and 2740 ROIs and category 2 having an AUC score of 85.74 and 2573 ROIs. Since the nipple of the breast is mostly located in the lower region of the breast, category 2 from 4 horizontal slices was more relevant in comparison to category 1. These instances were selected for further analysis.

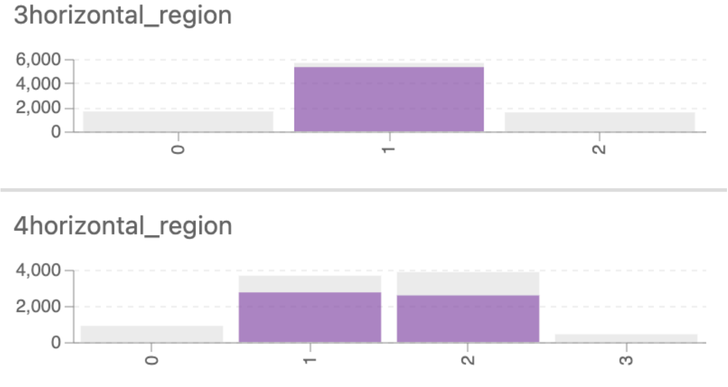


Figure 44 Correspondence between category 1 from 3 horizontal slices to categories 1 and 2 in 4 horizontal slices

Narrowing ROIs in category 2 from 4 horizontal slices further gives instances in categories 2 and 3 in 5 horizontal slices as in figure 45. Analyzing categories 2 and 3 in 5 horizontal slices showed that category 3 has a higher AUC score than category 2 of 87.36.

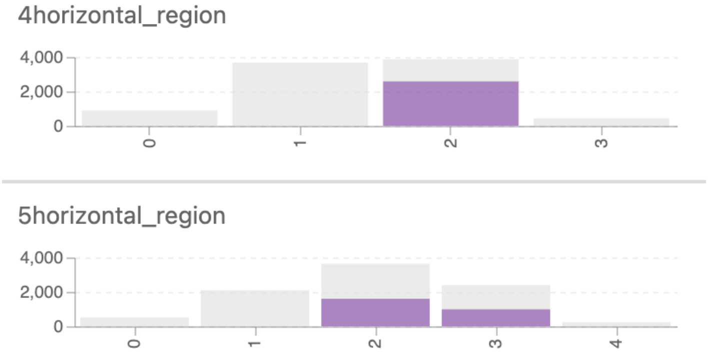


Figure 45 Correspondence between category 2 from 4 horizontal slices to categories 2 and 3 in 5 horizontal slices

Additionally, after analyzing mammograms, we found that category 3 in 5 horizontal slices corresponded to the area behind the nipple for most breast images. Therefore, this region was selected for the central breast region. When compared with all other categories, ROIs extracted from outside the area behind the nipple performed worse in comparison to ROIs taken from category 3 in 5 horizontal slices.

4.3 Analysis w.r.t. Horizontal and Vertical Position of ROI

Overall, considering the ROIs corresponding to categories 1 and 3 in 5 vertical slices and to category 3 in 5 horizontal slices, the AUC score we obtain is 89.17 and the total number of ROIs in this target area are 951. If we consider an ROI outside of the selected categories from the vertical slices or the selected category from the horizontal slices, then the AUC score decreases. These results are summarized in table 5.

	V	$\bar{V}$
H	AUC score: 89.17 No. of Instances: 951	AUC score: 87.56 No. of Instances: 2745
$\bar{H}$	AUC score: 87.81 No. of Instances: 1122	AUC score: 85.73 No. of Instances: 2980

Table 5 Set V corresponds to the selected categories from the 5 vertical slices and set H corresponds to the selected category from the 5 horizontal slices table, where $V = \{1, 3\}$ and $H = \{3\}$. Likewise, $\bar{V}$ corresponds to other categories from the 5 vertical slices and set $\bar{H}$ corresponds to the other categories from the 5 horizontal slices table, where $\bar{V} = \{0, 2, 4\}$ and $\bar{H} = \{0, 1, 2, 4\}$.

As represented by results in the table and as shown in previous studies, ROIs that have been extracted from the central breast region and from behind the nipple of the breast tend to have a better lesion prediction performance than ROIs extracted from other locations in the mammogram. Therefore, our current approach for extracting ROIs could be modified to take regions that might have more meaningful information to train the model. This holds particularly true for the ROIs extracted for BIRADS 1, that is, for the normal or no lesion mammograms because these ROIs are currently being arbitrarily extracted from the mammograms. In comparison to BIRADS 1, BIRADS 0 and 2 extract ROIs that have been specifically annotated by expert radiologists and as marked by the radiologists, these ROIs are the densest region(s) in that mammogram.

Chapter 5

Conclusion

Our interpretability experiments on the breast lesion detection model and the UPITT-mammogram dataset helped identify several patterns for analyzing the model performance. We found that although metadata and information derived from the data can be useful, it is still limited in terms of identifying the hidden information in the regions of interest. However, features analyzed in this step helped determine what aspects needed to be further explored. For instance, poor performance on normal patches with high contrast led towards investigating textural features in detail. Therefore, exploring and analyzing radiomic features for medical imaging can be better for identifying correlations between the model input and output. These correlations can be specifically relevant to help explain why the model makes a certain prediction on a specific output as well.

Our analysis of the relative position of a patch in the mammogram image was a confirming factor of the fact that if patches are extracted from the central breast region for model training, then the model can have better performance as the central breast region has most dense tissue. Therefore, we were able to identify another possible direction for further improving the performance of our model on normal patches. We found that patches that had been extracted from the central breast region performed better in comparison to patches that had been taken from outside the central breast region. Therefore, extracting patches from the central region of the breast for the normal mammograms during model training would be a better approach compared to arbitrarily extracting patches from the overall breast tissue region. Patches taken from the central breast region of normal mammograms would ensure that we have more dense and texturized images for the no lesion patches. Re-training the model on these patches could help generalize the model better and tackle the issue of the model's poor performance on coarse patches with high difference in neighborhood gray level intensities.

Limitations There are several limitations in our experiment and the dataset we worked with.

- The radiomic features we analyzed for our experiment were based on previous studies that had employed those features in their breast mammography studies. However, for our specific case, to best analyze what features would have been most important, it would have been better to adopt machine learning techniques for feature selection.

- There were some normal patches in our analysis that had a few problems. For instance, some patches had annotations. These annotations interfered with the radiomic feature values as they introduce texture that was not originally present in the patient's mammogram. Another problem with some normal patches was that since some of them were extracted from close to the breast boundary, these patches had plain black borders. These black regions interfered with first order radiomic features, such as mean.

- For analyzing the grid position of the ROIs in the mammogram, a major limitation was that given the current methods employed, it was difficult to determine the exact position of the ROI relative to the breast size and position in the original mammogram. This is because our dataset consisted of different left and right views, with different covered area of breast.

- There are some patches that are incorrectly labeled in the dataset as well. For instance, some patches with true labels as lesions consisted of no lesions.

Future Directions We propose several future works in the following directions:

- With the help of radiologists, it would be useful to visually group specific ranges of radiomic features that correspond to certain features in the breast. For instance, vessels could be marked by radiologists for specific radiomic feature values. This grouping of visual features can further improve interpretability of the model.

- Correctness of existing patches should be verified by radiologists so that all incorrect true labels can be fixed.

- To extract the accurate ROI position relative to the breast, we should augment all the mammograms in one direction, get a tight cropping around the full breast, expand the breast regions to be roughly the same, and then extract a central breast region patch from the pre-processed mammograms. This would eliminate the issue of the presence of black area/border in normal patches as well.

- Since we have calculated radiomic features, we can attempt to use these imaging characteristics for predictive models as well [1, 2]. This is because radiomic analysis has shown utility in predicting future risk of cancer as well [4].

- Currently, since the ROI is a square region, the shape radiomic features cannot be explored. In the future, it would be useful to investigate shape radiomic features given accurately masked lesion patches.

- To work with a model developer to test their model(s) by performing an analysis similar to our approach and determine if this helps them get insights on improving their model.

Bibliography

1. Lambin, P., Rios-Velazquez, E., Leijenaar, R., Carvalho, S., van Stiphout, R. G., Granton, P., Zegers, C. M., Gillies, R., Boellard, R., Dekker, A., & Aerts, H. J. (2012). Radiomics: extracting more information from medical images using advanced feature analysis. *European journal of cancer (Oxford, England : 1990)*, *48*(4), 441–446. https://doi.org/10.1016/j.ejca.2011.11.036

2. Lee, S. H., Park, H., & Ko, E. S. (2020). Radiomics in Breast Imaging from Techniques to Clinical Applications: A Review. *Korean journal of radiology*, *21*(7), 779–792. https://doi.org/10.3348/kjr.2019.0855

3. Limkin, E. J., Sun, R., Dercle, L., Zacharaki, E. I., Robert, C., Reuzé, S., Schernberg, A., Paragios, N., Deutsch, E., & Ferté, C. (2017). Promises and challenges for the implementation of computational medical imaging (radiomics) in oncology. *Annals of oncology : official journal of the European Society for Medical Oncology*, *28*(6), 1191–1206. https://doi.org/10.1093/annonc/mdx034

4. Mendel, K. R., Li, H., Lan, L., Chan, C. W., King, L. M., Tayob, N., Whitman, G., El-Zein, R., Bedrosian, I., & Giger, M. L. (2018). Temporal assessment of radiomic features on clinical mammography in a high-risk population. In K. Mori, & N. Petrick (Eds.), *Medical Imaging 2018: Computer-Aided Diagnosis* [105753Q] (Progress in Biomedical Optics and Imaging - Proceedings of SPIE; Vol. 10575). SPIE. https://doi.org/10.1117/12.2293368

5. Drukker, K., Giger, M. L., Joe, B. N., Kerlikowske, K., Greenwood, H., Drukteinis, J. S., Niell, B., Fan, B., Malkov, S., Avila, J., Kazemi, L., & Shepherd, J. (2019). Combined Benefit of Quantitative Three-Compartment Breast Image Analysis and Mammography Radiomics in the Classification of Breast Masses in a Clinical Data Set. *Radiology*, *290*(3), 621–628. https://doi.org/10.1148/radiol.2018180608

6. Li, H., Mendel, K. R., Lan, L., Sheth, D., & Giger, M. L. (2019). Digital Mammography in Breast Cancer: Additive Value of Radiomics of Breast Parenchyma. *Radiology*, 291(1), 15–20. https://doi.org/10.1148/radiol.2019181113

7. Karahaliou, A., Skiadopoulos, S., Boniatis, I., Sakellaropoulos, P., Likaki, E., Panayiotakis, G., & Costaridou, L. (2007). Texture analysis of tissue surrounding microcalcifications on mammograms for breast cancer diagnosis. *The British journal of radiology*, *80*(956), 648–656. https://doi.org/10.1259/bjr/30415751

8. Tagliafico, A. S., Valdora, F., Mariscotti, G., Durando, M., Nori, J., La Forgia, D., Rosenberg, I., Caumo, F., Gandolfo, N., Houssami, N., & Calabrese, M. (2018). An exploratory radiomics analysis on digital breast tomosynthesis in women with mammographically negative dense breasts. *Breast (Edinburgh, Scotland)*, *40*, 92–96. https://doi.org/10.1016/j.breast.2018.04.016

9. van Timmeren, J. E., Cester, D., Tanadini-Lang, S., Alkadhi, H., & Baessler, B. (2020). Radiomics in medical imaging-"how-to" guide and critical reflection. *Insights into imaging*, *11*(1), 91. https://doi.org/10.1186/s13244-020-00887-2

10. van Griethuysen, J., Fedorov, A., Parmar, C., Hosny, A., Aucoin, N., Narayan, V., Beets-Tan, R., Fillion-Robin, J. C., Pieper, S., & Aerts, H. (2017). Computational Radiomics System to Decode the Radiographic Phenotype. *Cancer research*, *77*(21), e104–e107. https://doi.org/10.1158/0008-5472.CAN-17-0339

11. Wang, G., Shi, D., Guo, Q., Zhang, H., Wang, S., & Ren, K. (2022). Radiomics Based on Digital Mammography Helps to Identify Mammographic Masses Suspicious for Cancer. *Frontiers in oncology*, *12*, 843436. https://doi.org/10.3389/fonc.2022.843436

12. Robinson, K., Li, H., Lan, L., Schacht, D., & Giger, M. (2019). Radiomics robustness assessment and classification evaluation: A two-stage method demonstrated on multivendor FFDM. *Medical physics*, *46*(5), 2145–2156. https://doi.org/10.1002/mp.13455

13. Singh, D., & Singh, M. (2016). Investigation on ROI Selection for Mammograms Using Texture Models and Machine Learning Classifiers. International Journal of Control Theory and Applications. 9.

14. Li, H., Giger, M. L., Huo, Z., Olopade, O. I., Lan, L., Weber, B. L., & Bonta, I. (2004). Computerized analysis of mammographic parenchymal patterns for assessing breast cancer risk: effect of ROI size and location. *Medical physics*, *31*(3), 549–555. https://doi.org/10.1118/1.1644514

15. Huo, Z., Giger, M. L., Wolverton, D. E., Zhong, W., Cumming, S., & Olopade, O. I., Computerized analysis of mammographic parenchymal pat- terns for breast cancer risk assessment: feature selection. *Medical physics*. 27, 4–12 (2000)

16. Rizzo, S., Botta, F., Raimondi, S., Origgi, D., Fanciullo, C., Morganti, A. G., & Bellomi, M. (2018). Radiomics: the facts and the challenges of image analysis. *European radiology experimental*, *2*(1), 36. https://doi.org/10.1186/s41747-018-0068-z

17. Contrast. Image contrast in Photoshop. (n.d.). Retrieved August 3, 2022, from https://helpx.adobe.com/photoshop/key-concepts/contrast.html#:~:text=The%20difference%20in%20brightness%20between,and%20dimension%2C%20and%20looks%20crisp

18. Paschali, M., Naeem, M.F., Simson, W., Steiger, K., Mollenhauer, M., & Navab, N. (2019). Deep Learning Under the Microscope: Improving the Interpretability of Medical Imaging Neural Networks. *ArXiv, abs/1904.03127*.

19. Huff, D. T., Weisman, A. J., & Jeraj, R. (2021). Interpretation and visualization techniques for deep learning models in medical imaging. *Physics in medicine and biology*, *66*(4), 04TR01. https://doi.org/10.1088/1361-6560/abcd17

20. Timothy B. Lee (2019). How neural networks work-and why they've become a big business. Retrieved August 3, 2022, from https://arstechnica.com/science/2019/12/how-neural-networks-work-and-why-theyve-become-a-big-business/

21. Singh, A., Sengupta, S., & Lakshminarayanan, V. (2020). Explainable Deep Learning Models in Medical Image Analysis. *Journal of imaging*, *6*(6), 52. https://doi.org/10.3390/jimaging6060052

22. Reyes, M., Meier, R., Pereira, S., Silva, C. A., Dahlweid, F. M., von Tengg-Kobligk, H., Summers, R. M., & Wiest, R. (2020). On the Interpretability of Artificial Intelligence in Radiology: Challenges and Opportunities. *Radiology. Artificial intelligence*, *2*(3), e190043. https://doi.org/10.1148/ryai.2020190043

23. Salahuddin, Z., Woodruff, H., Chatterjee, A., & Lambin, P. (2021). Transparency of Deep Neural Networks for Medical Image Analysis: A Review of Interpretability Methods.

24. Lévy, D., Jain, A. Breast mass classification from mammograms using deep convolutional neural networks. arXiv 2016, arXiv:1612.00542

25. Lee, H., Kim, S.T., Ro, Y.M. Generation of Multimodal Justification Using Visual Word Constraint Model for Explainable Computer-Aided Diagnosis. In Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support; Springer: Cham, Switzerland, 2019; pp. 21–29.

26. Yeche, H., Harrison, J., & Berthier, T. (2019). UBS: A dimension-agnostic metric for concept vector interpretability applied to radiomics. In *Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support* (pp. 12-20). Springer, Cham.

27. S. Li, M. Dong, G. Du and X. Mu, "Attention Dense-U-Net for Automatic Breast Mass Segmentation in Digital Mammogram," in IEEE Access, vol. 7, pp. 59037-59047, 2019, doi: 10.1109/ACCESS.2019.2914873.

28. Shen, L., Margolies, L.R., Rothstein, J.H. *et al.* Deep Learning to Improve Breast Cancer Detection on Screening Mammography. *Sci Rep* **9,** 12495 (2019). https://doi.org/10.1038/s41598-019-48995-4

29. PyRadiomics Documentation. Retrieved August 3, 2022, from https://pyradiomics.readthedocs.io/en/latest/

30. Tietz, E., Truhn, D., Müller-Franzes, G., Berres, M. L., Hamesch, K., Lang, S. A., Kuhl, C. K., Bruners, P., & Schulze-Hagen, M. (2021). A Radiomics Approach to Predict the Emergence of New Hepatocellular Carcinoma in Computed Tomography for High-Risk Patients with Liver Cirrhosis. *Diagnostics (Basel, Switzerland)*, *11*(9), 1650. https://doi.org/10.3390/diagnostics11091650