Computer-Aided Diagnosis (CAD) of Breast Cancer: Methods of Model Explainability

Teresa M. Bodart

Data Science, University of Wisconsin - La Crosse

DS 785: Capstone

Dr. Tracy Bibelnieks

December 11, 2022

Abstract

The International Agency for Research on Cancer announced that in 2020 female breast cancer became the most diagnosed cancer worldwide and the most common cause of cancer-related death in women. Still, breast cancer generally has a good prognosis with timely detection and appropriate treatment. Recently, computer-aided diagnosis (CAD) systems have shown promising results in using artificial intelligence (AI) to detect malignant lesions in breast ultrasound (US) imaging. When working with AI in a clinical setting, however, the American College of Radiology advocates for radiologist understanding of the algorithms in use. Accordingly, this study contributes to an ongoing collaboration between the University of Wisconsin-La Crosse and Mayo Clinic Enterprise by investigating three methods of AI explainability for the CAD software in development. Class activation maps, saliency maps, and attention map-enhanced class activation maps were compared to determine the most useful technique for visualizing regions in the US used by the models to determine pathology. The evaluation showed that saliency maps are the most promising method for visually explaining breast US classification. However, the small dataset and simplified model architecture used in this study mean that further research is necessary before fully implementing this method within the greater collaboration.

Keywords: Breast cancer, ultrasound, radiology, deep learning, explainability

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Background

Among women in the United States, breast cancer is the second most frequent cancer diagnosis and cause of cancer-related death, leading to an estimated 287,850 new cases of invasive breast cancer and 43,250 deaths in 2022 (Giaquinto et al., 2022). In 2020, female breast cancer became the most commonly diagnosed cancer in the world and the most common cause of cancer-related death in women (International Agency for Research on Cancer, 2021). These figures and trends highlight the magnitude of the situation and underscore the importance of continued research in the field of breast cancer.

The essential components of effective breast cancer diagnosis and treatment are clinical breast assessment and examination followed by diagnostic imaging, tissue sampling, prompt surgery, and systemic therapy (Ginsburg et al., 2020). Diagnostic imaging presents an opportunity for the practical application of data science because the images are rich in clinical indicators not always visible to the human eye (Agarwal, 2022). This project focuses on the diagnostic imaging stage.

In the case of breast cancer, mammography and breast ultrasound (US) comprise two of the most used imaging modalities. According to Shen, Shamout, et al. (2021), while mammography is the most widely used technique for imaging, it has significant limitations: it is not always accessible and can produce poor results in patients with dense breast tissue; therefore, breast US often serves as a supplementary modality in screening and as the primary modality in diagnosis. Further, breast US has been found to have a comparable cancer detection rate to that of mammography (Berg et al., 2015).

This project will contribute to an ongoing collaboration between the University of Wisconsin-La Crosse and Mayo Clinic Enterprise (BUS Project). The initiative aims to develop a state-of-the-art Computer-Aided Diagnosis (CAD) system for breast US lesion interpretation by using large case studies from Mayo Clinic Enterprise and advanced AI technology supported by UWL Mathematicians, Data Scientists, and graduate students.

Statement of the Problem

Breast US is an important modality for the screening and diagnosis of breast cancer. However, like mammography, this technique is not without drawbacks and limitations. In the same study that found breast US to have a comparable detection ability to mammography, the authors also highlighted the higher false-positive rate (i.e. the rate that patients without cancer are called for further screening or biopsy) of breast US (Berg et al., 2015). Across Mayo Clinic Enterprise, the positive biopsy rate following a breast US ranged from 31-51% in 2019 and 2020. These biopsies following a false-positive breast US classification are unnecessary procedures and place undue stress on patients and their families, as well as on the healthcare system.

In addition, ultrasound is very user-dependent as an imaging modality and has the largest range of radiologist lesion interpretation and positive biopsy rates. Evans et al. (2018) recommend that women referred for breast US be informed that the operator's specific competence and experience can dramatically influence the diagnostic ability of their ultrasound. This variation creates non-uniform patient care across the health system and can exacerbate already present disparities.

Recent advances in artificial intelligence (AI) and deep learning, particularly convolutional neural networks (CNNs), have the potential to address some of these limitations when applied to image classification and object detection. However, the application of deep

learning and AI to sensitive health data within the medical field presents its own challenges and limitations. A deep learning neural network such as a CNN has hidden layers that can be non-intuitive and offer no insights into how or why the algorithm makes intermediate decisions. Despite this, some level of explainability must necessarily be built into the product of this project to achieve a level of trust and transparency for use by radiologists in breast cancer diagnosis.

Conceptual Framework

The partnership between Mayo Clinic and the University of Wisconsin-La Crosse is structured into three phases:

1. Build and train deep learning models to classify breast US lesions from high-quality Mayo Clinic data into the correct Breast Imaging Reporting & Data System (BI-RADS®) Assessment Category. BI-RADS is a system used by radiologists for lesion assessment and to clarify communication with physicians and patients (Mendelson et al., 2013). This approach is completely automated and functions as a black box.
2. Build a second algorithm to classify lesions that mimics systems used by experts. As these systems rely on characteristics of the lesion normally determined by a radiologist—such as orientation, margins, elasticity, and vascularity—machine learning algorithms and rule-based algorithms will be used alongside optional user input.
3. Combine the models developed in phases one and two to produce a single system that outperforms each of the individual models.

At present, the BUS Project is in its second year and collaborators are currently working on all three objectives; in particular, another graduate student is making significant progress in applying state-of-the-art segmentation models for more accurate lesion segmentation. These results can be used to compute features and characteristics for the machine learning algorithms.

The project described herein will focus on the black box aspect of algorithms created according to the first approach. While it is true that artificial neural networks were inspired by the biological neural network structure of the brain, the internal workings and training processes are poorly understood—and without understanding how an algorithm achieved certain results, it is difficult to then trust those results (Cole, 2020). Still, methods are being developed to introduce explainability into some deep neural network algorithms.

Working with breast US image data, this project implements and evaluates three different techniques that create visual explanations for breast US classification by deep learning algorithms: (a) class activation maps, (b) saliency maps, and (c) attention map-enhanced class activation maps. Each method interacts uniquely with the algorithm and generates visualizations highlighting parts of the US that were important for classification. However, attention map-enhanced class activation maps are implemented using a vision transformer model architecture, which has only recently begun to be studied on medical imaging. Therefore, the evaluation of this method is preliminary in the context of the BUS Project. In addition, because the central concept of this study explores the feasibility of each visualization technique, the deep learning algorithms utilized will output binary lesion classifications—benign or malignant—rather than the more complex BI-RADS classification system.

Purpose of the Study

Usefulness and practical application of any CAD software must stay at the forefront of any study aiming to bring cutting-edge technologies to clinical practice. With AI algorithms applied to medical imaging, this is especially important. The American College of Radiology advocates for physician understanding—at a high level how algorithms are developed and

trained, how they perform, how they should be used, and their limitations—when working with AI in a clinical setting (Sendak et al., 2021).

The purpose of this project is to support the goal of practical application by building algorithm explainability into the CAD framework. Accordingly, three methods of visual explanation for AI classification of breast US imaging will be evaluated to determine the most useful technique for radiologists and technicians who conduct imaging. In the case of this collaboration between Mayo Clinic and the University of Wisconsin-La Crosse, there is a unique opportunity to receive direct input and feedback from a practicing radiologist. This feedback will aid in the evaluation of user experience and usefulness.

Significance of the Study

The application of deep neural networks to medical imaging is a fast-moving field that continues to develop. Studies have found that state-of-the-art CAD algorithms using deep learning for binary classification of breast US lesions have achieved as high as 87% accuracy (Cao et al., 2019). Other benefits of AI applications to breast ultrasonography include greater workflow efficiency and a reduction of interobserver variability—the variability between radiologists that contributes to non-uniform care (Kim et al., 2021). Altogether, the responsible implementation of deep learning in breast US lesion diagnostics could have tremendous benefits.

In particular, this collaboration represents an integral step forward in the responsible implementation of AI on breast US imaging. If the high-false positive rate of breast US can be mitigated, and a CAD system that promotes transparency and trust can be designed, it is the hope of this study to improve patient care for individuals and benefit the health system as a whole.

Organization of the Project

The following objectives outline the project structure and how it will meaningfully contribute to the BUS Project overall:

1. Compile breast US images from multiple sources for training Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs).
2. Apply deep learning algorithms to Mayo Clinic breast US images and build a model with 80-90% accuracy in classifying lesions as benign or malignant.
3. Implement class activation maps (CAMs), saliency maps, and attention map-enhanced CAMs with deep learning algorithms on Mayo Clinic breast US images to establish interpretability in the BUS Project and increase usability by radiologists.
4. Present outcomes from various methods of model explainability through appropriate data visualizations and evaluate with project team and client to determine the most useful method to aid radiologists in lesion interpretation and incorporate the selected method into the project codebase.

Limitations of the Study

Deep learning algorithms often require extremely large amounts of data for training and validation to achieve a high level of accuracy. Currently, the number of high-quality, labeled breast US images available from the Mayo Clinic is limited. Efforts are being made to increase the size of the dataset and find solutions to the lack of data issue, but the algorithms' performance are not guaranteed to reach levels found in similar studies. While the focus of the project described here is on evaluating methods for algorithm explainability, poor model performance could negatively affect the results. In turn, there could be difficulty in determining the most appropriate technique.

Additionally, it is worth noting again that the full CAD software, once completed, will be making assessments according to the BI-RADS system for evaluating breast US. Because the scope of this study is centered around visualization techniques, the deep learning models utilized will only be making binary classifications between benign and malignant.

Chapter 2: Literature Review

Introduction

The application of deep learning algorithms to medical data is an evolving field that shows great promise. In particular, diagnostic imaging presents a golden opportunity to apply sophisticated algorithms and achieve results similar to—or better than—those achieved by a trained radiologist. The collaboration between the University of Wisconsin-La Crosse and Mayo Clinic aims to apply deep learning algorithms to high-quality breast US images provided by Mayo Clinic, eventually creating a state-of-the-art CAD software to aid radiologists in diagnosing breast cancer lesions, thus reducing the high false-positive rate and ensuing unnecessary biopsies attributed to breast US as a diagnostic modality.

This project deals specifically with methodologies for explaining how deep learning algorithms arrive at classification results. Three main techniques are used to output visualizations depicting the discriminative image regions important to classification. However, some of these methods have not been rigorously studied and applied to breast US images, while others have served a different purpose than model explainability. This literature review will cover the various deep learning model architectures utilized for medical image classification and their most relevant interpretability techniques. A review of their respective applicability to medical imaging and breast US will also be presented.

Deep Learning Computer-Aided Diagnosis Systems

Over the past decade, computer-aided diagnosis (CAD) systems in the medical field have advanced alongside the rapid development of deep learning. Jiménez-Gaona et al. (2020) present a critical review of how deep learning has been applied to breast tumor diagnosis from 2010-2020. The authors found that convolutional neural networks (CNNs) are the most popular models used for intelligent image analysis and cancer detection through their review, leading to the prominence of Deep Learning (DL) CAD systems in recent studies. While providing good performance, CNNs also automate feature extraction, which has historically been a manual task used as input for traditional Machine Learning (ML) CAD systems (Cao et al., 2019). Among the most commonly used metrics for evaluating DL-CAD system performance are accuracy, sensitivity, specificity, and confusion matrices, all of which will be measured in this study. Overall, Jiménez-Gaona et al. concluded that DL-CAD can achieve better performance than traditional ML-CAD, although there is still a need to improve the models, specifically with larger and more balanced data sets and better optimization methods. Data scarcity and class-imbalances were two obstacles in this study, and they continue to pose major challenges to the BUS Project. The paper concludes by emphasizing the importance of CNNs in DL-CAD systems and laying out some of the challenges that must still be overcome.

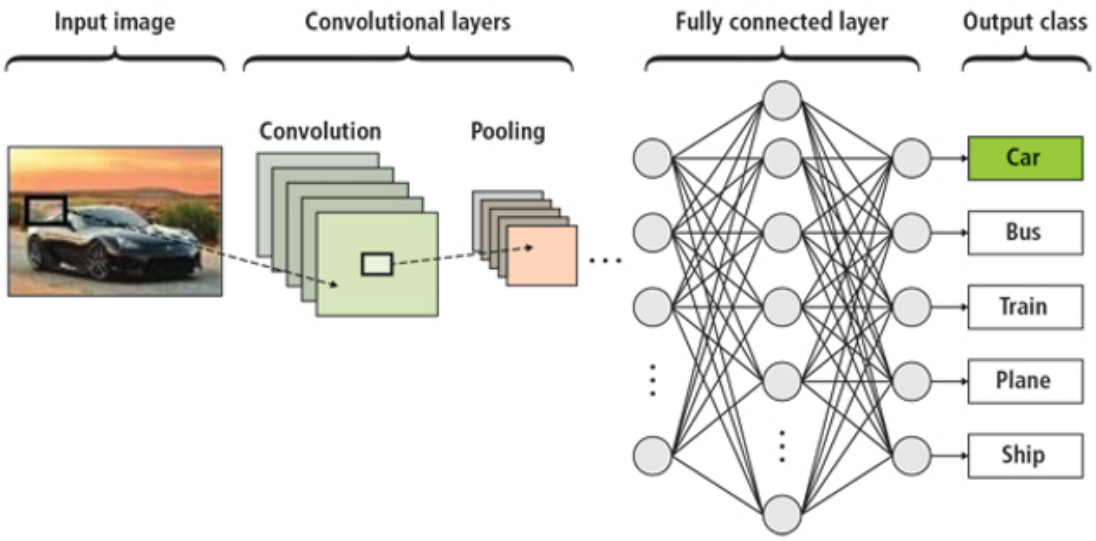
Convolutional Neural Networks

As noted above, CNNs are the most common model used for computer vision tasks in breast cancer diagnostic imaging. Computer vision is a field of artificial intelligence (AI) where computers extract meaningful information from images, which has applications in many fields including healthcare (IBM Cloud Education, 2020). The areas of medical diagnosis where computer vision and CNNs can be used are numerous, but the basic building blocks are often

similar. Figure 1 illustrates the three layers of CNNs: the input layer, the hidden layer consisting of convolutions, pooling, and normalization, and the output layer.

Figure 1

The Building Blocks of CNN Models (NVIDIA, n.d.)



First, the input image goes through the convolutional layers where filters are repeatedly applied to shifting sections of pixels, eventually producing a feature map that indicates the locations and strengths of detected features (NVIDIA, n.d.). Different filters might detect vertical lines, horizontal lines, edges, or degrees of light intensity. Compiling these outputs can reveal complex objects or elements from the training data. Next is the pooling layer which reduces the size of the feature maps to save on computation, often taking either the maximum values only or the average values. Normalization then occurs, stabilizing the network. Finally, the fully connected layer connects neutrons between layers and produces an output.

Yu et al. (2021) explain the inner workings of the most popular CNN models in use and expand upon their performance in medical imaging for different organs and conditions. The authors attribute the success of CNN in medical imaging to its built-in data pre-processing

techniques such as image normalization and augmentation, and to hyper-parameter optimization for parameters such as learning rate. Among the models described—AlexNet, GoogleNet, Deep Residual Network (ResNet), Regions with CNN Features (R-CNN), and Fully convolutional neural networks (F-CNN)—the authors noted that ResNet keeps more of the input information intact and does not produce relatively higher training error in deeper networks than other models. It does this by creating shortcut connections between layers, avoiding degradation that can occur due to repeated multiplication. A previous study within the BUS Project found the ResNet architecture to be the most suited for breast US out of the common models listed above (Jarvey, 2022).

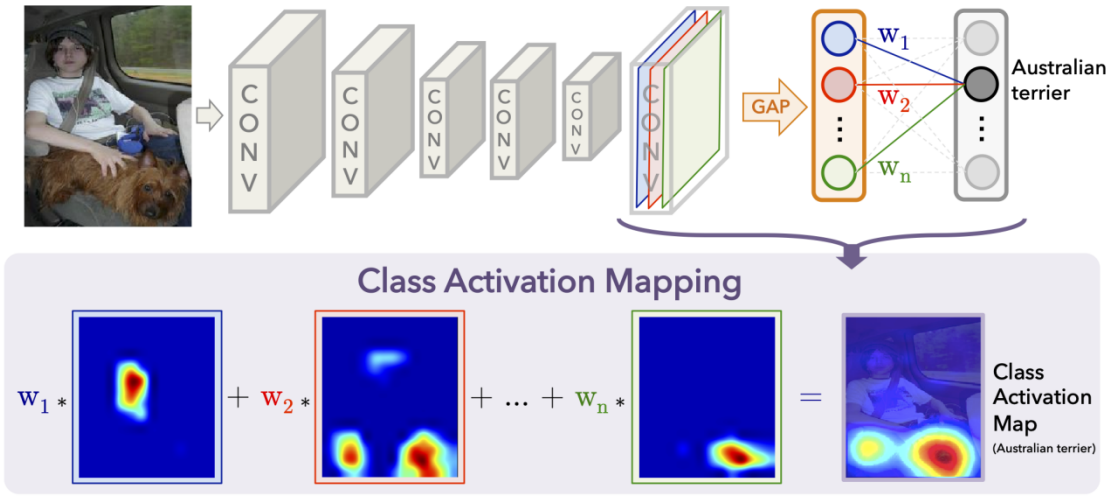
Class Activation Maps

One of the originating methods for visualizing why or how CNNs make predictions are class activation maps (CAMs), which are heatmaps produced by a model that highlight regions of an image that were important to classification. CAMs have gone on to precipitate increasingly advanced and specialized methods such as Gradient-weighted (Grad) CAM, GradCAM++, HiResCam, and EigenCAM. The foundation of CAMs is that some CNN layers already perform object detection using only class-level labels (Zhou et al., 2015). Zhou et al. (2016) demonstrate that by replacing the fully-connected layers before model output with pooling—global average pooling (GAP) or global max pooling (GMP)—and a fully-connected softmax layer, CNN classification performance is not significantly impacted and a CAM can be created showing the discriminative parts of the object detected. Figure 2 illustrates how the class-specific activations are weighted by relevance to the predicted class and then pooled to map the regions showing an Australian terrier, in this case. The authors also note the distinction between GAP and GMP:

GMP encourages the algorithm to highlight only the maximum discriminative region, thereby potentially overlooking other important regions.

Figure 2

Class Activation Mapping using Global-Average Pooling (Zhou et al., 2015)



In application to automated diagnosis of breast US, Qi et al. (2019) found that using CAMs as additional input to their model to enhance the region used to classify lesions experimentally resulted in improvements to model performance. In this case, CAMs were not used as an interpretability aid but still demonstrated their value to DL-CAD systems for breast US diagnosis.

While not yet studied in application to DL-CAD of breast US, two other descendants of CAM—Grad-CAM and HiResCAM—introduced important breakthroughs in the field of AI explainability. Grad-CAM, proposed by Selvaraju et al. (2019), does not suffer from the same tradeoff between model complexity and model transparency as CAM. The authors show that it is possible to generate class-discriminative visual explanations for any CNN without modifying the architecture of the model as is done with CAM. In addition, the authors define two criteria for a

good visual explanation of CNN: (a) localize the category or class in the image, and (b) have a high resolution for capturing fine details.

HiResCAM, on the other hand, confronts a recently discovered limitation of Grad-CAM in that it can sometimes highlight regions of an image that were not actually used by the model. Draelos & Carin (2020) propose HiResCAM as a generalization of CAM—it works on all CNNs—that guarantees faithful explanations of CNNs. As trust and transparency are key aims of this study, ensuring that visual explanations of an algorithm's predictions are faithful is a relevant consideration. For example, a physician needs to know if an algorithm actually used the region of the breast lesion to make a classification. This is an important aspect of the responsible application of algorithms to healthcare.

Saliency Maps

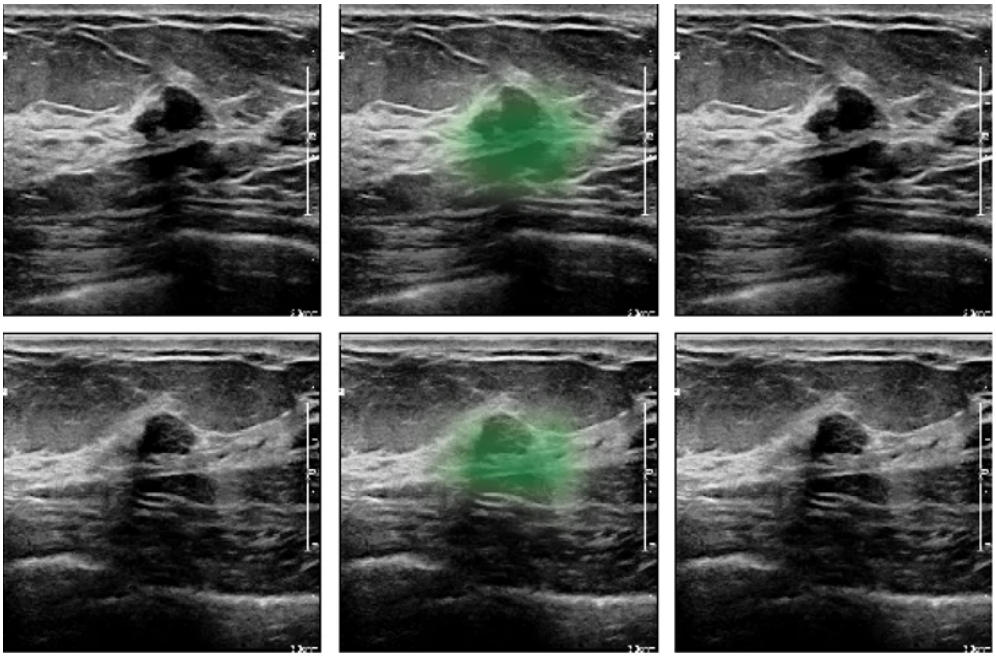
Saliency maps are another important methodology for visualizing and interpreting how a deep learning model achieves a result. This technique, first presented by Simonyan et al. (2014), can work with any CNN classification model to derive a class saliency value for every pixel in an image. These values are then composed into a saliency map depicting the relative importance of each pixel for the selected class. Superimposed upon the original image, a heatmap takes shape showing the salient regions of the image used for classification. Simonyan et al. also note that this method can be extremely fast—only one backpropagation pass required—and can be achieved using only image labels and no bounding boxes or segmentation masks.

Two studies using deep learning and advanced AI have had great success in accurately classifying breast mammography and ultrasound images as benign or malignant by utilizing saliency maps to narrow the region of interest (ROI) input into the models. Shen, Wu, et al. (2021) proposed a novel framework where a global module accepts an entire mammograph

image and outputs a pixel-level saliency map highlighting the ROIs: regions the models believe contain a malignant lesion. Then the image is cropped to focus on the ROI and passed into the local module, where a final classification is made. In essence, the saliency map segments the lesion and improves the performance of the final classification model. The researchers found that this model design surpassed radiologist-level diagnostic abilities in a reader study.

Figure 3

Saliency Map of Breast Ultrasound (Shen, Shamout, et al., 2021)



Shen, Shamout, et al. (2021) applied a similar model framework to breast US images and similarly achieved a higher area under the receiver operating characteristic curve (AUROC) than the average AUROC of ten board-certified breast radiologists surveyed. In addition, the study found that with the assistance of the AI model, radiologists were able to decrease their false-positive rate and reduce requested biopsies. Lastly, this second study placed greater importance on model explainability than the previous study, highlighting how their saliency maps (see

Figure 3) develop clinician trust and understanding, especially of the potential limitations of the model.

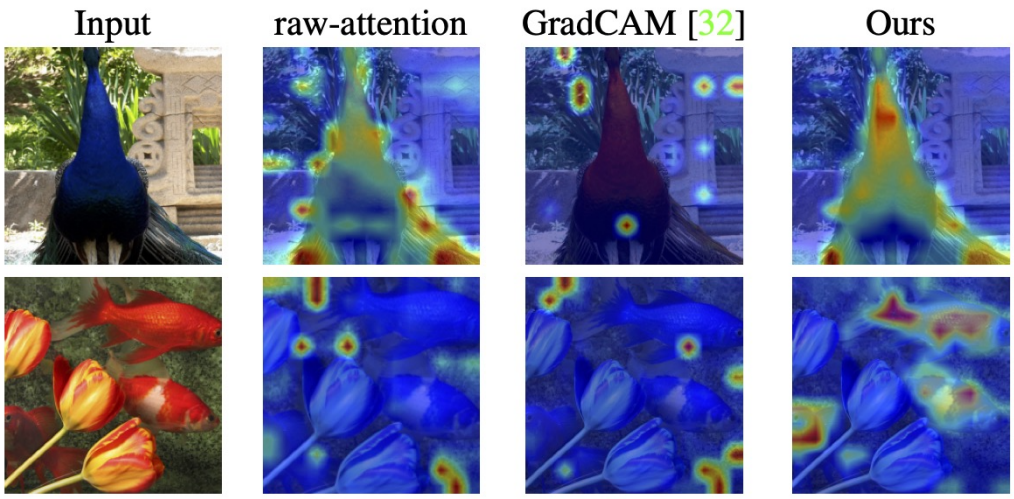
Vision Transformers and Conformers

In contrast to the previously described methodologies for explaining CNN models—CAMs and saliency maps—it was recently proposed that CNNs can be replaced by vision transformers (ViTs) in medical diagnostic imaging. Matsoukas et al. (2021) explored whether the transition from CNNs to ViTs could constitute a trivial change in practice but result in equal or better performance alongside additional advantages. The findings in their study applied to three mainstream healthcare datasets, including a mammography dataset, show that with sophisticated pre-training methods ViTs outperform their counterpart CNNs. Additionally, the transition was essentially seamless while gaining improved model explainability due to the built-in attention maps created by ViTs. The authors also note that ViT performance suffers to a greater degree than CNN in situations where data is scarce. This emphasizes the need for rigorous model pre-training.

At the core of ViT networks are are self-attention layers where pairwise attention values are assigned between patches of the image (Dosovitskiy et al., 2021). To visualize the model, these attentions are often considered as relevance scores and the results from a single layer are used as a heatmap (Chefer et al., 2021). Chefer et al. propose a framework that combines attentions from multiple layers in a way that maintains the sum of the relevancy. The remarkable performance of this framework can be seen in Figure 4 comparing many popular AI explainability techniques.

Figure 4

Attention Map Comparison (Chefer et al., 2021)

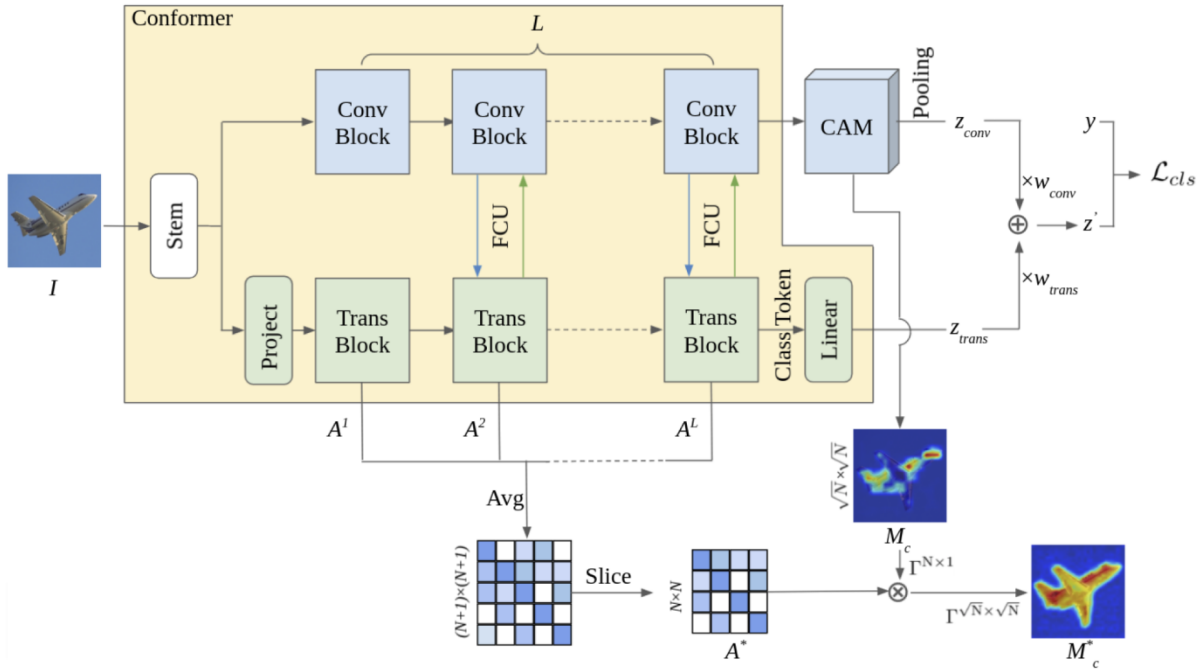


Taking attention maps one step further, a model architecture that utilizes a CNN branch and ViT branch was introduced by Peng et al. (2021), called a Conformer. This hybrid network structure takes advantage of a CNN's ability to capture local features, such as lines, edges, and shapes, and a ViT's ability to capture long-distance feature dependencies, or global representations. The focus of their study was on demonstrating the viability of the Conformer as a general backbone network, rather than the application to any specific computer vision task.

Built on a Conformer backbone, Li et al. (2022) then proposed TransCAM, a method using only image-level labels for refining the CAMs generated by a CNN with the attention maps produced by a ViT. The model framework can be visualized in Figure 5. This method for implementing a ViT and generating a visual explanation is fitting for this study, given that only image-level labels were used to train the models. Lastly, as TransCAM is such a newly proposed architecture, it has not yet been studied in application to medical imaging.

Figure 5

TransCAM Framework (Li et al., 2022)



Note. Input image *I* of a plane is seen on the left, with the enhanced CAM M^*_c generated by TransCAM visualized on the right. The dual-branch Conformer network with *L* convolutional and transformer blocks can be seen in the yellow box. Attention map A^* is constructed by averaging the self-attention weights from the transformer blocks. The CAM M_c generated by the CNN branch is refined by the attention map A^* through a single-step dot product refinement, resulting in the improved CAM M^*_c.

Conclusion

The various deep learning model architectures and methodologies for explaining AI have been discussed and their recent application to medical imaging and breast US diagnosis presented. A common theme in the literature is the multipurpose usage of the model explainability techniques, both to introduce transparency and interpretability and to act as functioning parts of the AI architecture. Class activation maps and saliency maps have both been

used by CAD systems as additional inputs to help the model locate the lesion and then make a classification. The studies using saliency maps to determine ROIs are perhaps the most similar to this project, albeit with different primary purposes for the visualizations. The focus of this study will principally be on the value of these methods as additional transparency for a radiologist using DL-CAD for breast US diagnosis, rather than as an input to the actual CAD software. Training the deep learning algorithms—basic CNN, Saliency-adapted CNN, and TransCAM—is a stepping stone to allow the creation of model explainability visualizations when the algorithms are specifically applied to breast US images. The comparison and evaluation of the methods will be qualitative rather than quantitative, focusing on how a radiologist might view the visualization and what information is communicated.

Chapter 3: Methodology

Introduction

The CAD software produced by the BUS Project collaboration between the University of Wisconsin-La Crosse and Mayo Clinic aims to match or exceed the breast US diagnostic accuracy achieved by board-certified radiologists. While the software has not reached its final iteration, by design it utilizes sophisticated deep learning algorithms to address the computer vision task. The software will also provide supplemental information to users, thereby becoming a valuable tool for radiologists and helping to reduce false-positive diagnoses and unnecessary biopsies.

Model explainability will be an important piece of the supplemental information. The method of model explainability should provide a radiologist with insight into how the model determined its output and confidence in the model's prediction, allowing them to responsibly apply the software. However, deep learning algorithms are by nature incredibly complex and

difficult to interpret. Prompted by this challenge, the research presented here implemented and evaluated three methods of model explainability for deep learning algorithms applied to breast US images.

The current standard model architecture for CAD on medical imaging is CNN, for which two methods of model explainability were evaluated. One preliminary method of Conformer—combined ViT and CNN branches—model explainability was also investigated as transformers have recently been found to achieve similar or better results compared to CNNs when applied to generic image datasets. However, the application of transformers to medical imaging is in its infancy, meaning that the analysis of this method was only exploratory and not ready for integration with the larger BUS Project.

Concerning the data for this project, large datasets are often necessary to produce desired results from deep learning algorithms. While collaborators at Mayo Clinic are working to anonymize and annotate more high-quality breast US images, the number of images remains small. For this reason, public breast US datasets supplemented the data for this project to aid in model performance.

Data Collection

The Mayo breast US dataset used in this study was provided by Mayo Clinic and contained 275 images from 153 unique patient studies as of November 25, 2022. For many patients, an image of the lesion was taken from both the longitudinal and transversal probe angles—this meant that the dataset contained multiple images of some lesions. The average image resolution was 560 x 560 pixels in height and width, respectively. A binary pathology result of either benign or malignant was reported for each image, along with radiologist annotations on lesion characteristics and other patient information. The Mayo Clinic's data

privacy team handled data anonymization to ensure all HIPAA and privacy regulations were met, removing all personally identifiable information (PII) from each image. To keep track of all relevant information related to the images, randomly generated external IDs were used to replace patient identification numbers. The images were then saved in PNG format with the naming convention 'externalID_pathology_orientation.png'.

To augment the small number of images available through Mayo Clinic, three public breast US datasets were added to the project data for model training. Table 1 summarizes the contributions and relevant features of each dataset. The first, Breast Ultrasound Dataset B (BUS_Dataset_B), gave no indication of whether any images were from the same patient (Yap et al., 2017). The second public dataset, Breast Ultrasound Images with Ground Truth (Dataset_BUSI_with_GT), contained "normal" images in addition to benign and malignant (Al-Dhabyani et al., 2020). The normal images did not contain any lesions and were filtered out for this study. Additionally, the source did not provide information on which images corresponded to the same patient. The last supplemental data was Breast Ultrasound Videos (BUV), with 75 videos of benign lesions and 113 videos of malignant lesions (Lin et al., 2022). The individual frames from each video were saved as PNG files. Because adjacent frames were nearly identical, each case was limited to include every 50th frame.

In total, 1678 breast US images were collected to train, test, and visualize deep learning algorithms. The dataset source, image file path, and pathology of each image were collated into one dataframe. The dataframe also maintained the connection between images from the same patient or lesion. Minor data cleaning was necessary to remove two duplicate images from the Mayo dataset and filter the normal US images that did not contain lesions.

Table 1

Data Sources

Dataset Name	Number of Images			Unique Patients	Average Resolution (hxw)
	Total	Benign	Malignant		
Mayo	275	144	131	153	560 x 560
BUS_Dataset_B	163	109	54	—	450 x 540
Dataset_BUSI_with_GT	647	437	210	600	490 x 610
BUV	593	233	360	188	620 x 620
Total	1678	923	755	—	550 x 600

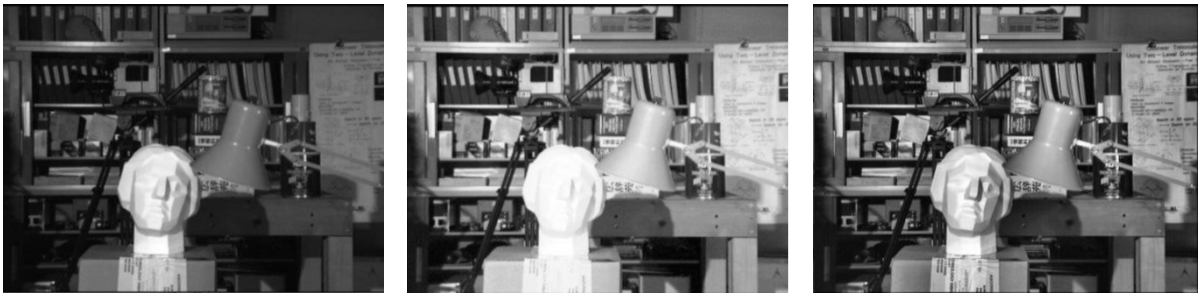
Data Preparation

Breast US images are captured in grayscale, which can make fine details difficult to distinguish. Histogram equalization is a method of image processing that can increase the contrast in areas of an image with low contrast and expose features that could be useful for image recognition and classification tasks. To prepare the data for modeling, Contrast Limited Adaptive Histogram Equalization (CLAHE) was applied to each image. CLAHE was used instead of global histogram equalization because details can be lost when the global contrast is used. Instead, CLAHE applies histogram equalization to small sections of an image. Figure 6 illustrates the detail that can be uncovered or suppressed depending on technique.

After all images in the dataset were processed through OpenCV's CLAHE function they were stored in a new column of the original dataframe as two-dimensional NumPy arrays. This step also converted all images to the one-channel grayscale format even though some were originally saved in three-channel RGB format.

Figure 6

Global Histogram Equalization Compared to CLAHE (OpenCV, n.d.)



(a) Original image | *(b) After global histogram equalization* | *(c) After CLAHE*

Note. Compare details on the statue's face: (b) causes a loss of information due to over-brightness, but (c) shows improved contrast throughout the image and more details on the statue's face.

The final step of data preparation was splitting the images into train and test sets. A separate model validation set was not necessary because the focus of this study was on methods for extracting model explainability visualizations from deep learning algorithms and not on model optimization or fine-tuning. However, it was important to account for the numerous patients with multiple breast US images, the variability in quality between source datasets, and the distribution of benign and malignant images.

Consequently, patients were stratified by dataset and pathology and then randomly split among training (80%) and test (20%) sets, totaling 1347 training images and 331 test images. This method of splitting ensured that each patient or lesion only appeared in one of the training and test sets, and the model was not tested on a lesion it had been trained on. In addition, the stratification method maintained the relative distribution of datasets and pathologies between training and test sets.

Modeling

This section details the process of building and training simplified versions of the deep learning algorithms used by the BUS Project. The simplifications included using a binary pathology category rather than the numeric BI-RADS score that the final CAD software will produce, performing only basic data augmentations, and forgoing optimization. Together, these adjustments allowed more time to be dedicated to the methods of model explainability described in the next section.

Data Augmentation

Data augmentation is a technique that can be used to artificially increase the amount of data in a dataset by generating new data points from the existing data. This improves the performance of deep learning algorithms and mitigates the challenge of insufficient data. Two geometric augmentations were applied to the data during training to improve model performance. The two transformations used were FlipItem and RandomResizedCrop, both studied and implemented by Jarvey (2022), who worked on model optimization for the BUS Project. FlipItem randomly flipped images in the dataset along their vertical axes with a probability of 0.50, creating a new image to train the model on. Images could not be flipped over the horizontal axis because the model should never see an image in the upside-down orientation, given the standard method of US capture. RandomResizedCrop took a random scaled crop of each image containing at least 80% of the area of the original image and resized the crop to 256 x 256 pixels. This was necessary to fit the input requirements of the pre-trained ResNet-34 backbone model. Figure 6 shows an example of a histogram equalized breast US image after data augmentations and illustrates the data used for model training. Other augmentation techniques such as rotation, zoom, brightness, and warp were considered, but the marginal effects on model performance were found to be negligible and therefore these augmentations were left out.

Figure 7

Histogram Equalization and Data Augmentations

Convolutional Neural Networks

The model architectures used for the two methods of CNN explainability—CAMs and saliency maps—were very similar: both followed the basic CNN structure described in Chapter 2, with convolutional layers, pooling layers, normalization, and a fully-connected layer that output classification. Additionally, both used the state-of-the-art backbone network ResNet-34. This is a model that was pre-trained on generic images from ImageNet and then fine-tuned for the new image processing task (Huh, 2016). While there are multiple popular backbone networks, such as VGG-16 and DenseNet-201, ResNet-34 was selected because a previous BUS Project capstone found it to have the best performance when applied to Mayo Clinic breast US images (Jarvey, 2022).

However, a few key modifications were made to the model architecture to generate saliency maps. As noted in Chapter 2, CAMs can be produced without altering the model architecture in any way, whereas it is necessary to alter the basic CNN model for saliency maps. These changes follow the first portion of the model framework used by Shen, Shamout, et al. (2021) to produce saliency maps. After the base convolutional filters have been applied, a

sigmoid was applied to the weighted average of the feature maps to produce a saliency map for each possible classification: benign and malignant. These maps spotlight approximate regions of benign or malignant lesions in each image while recording the contribution of each location to the pathology prediction.

The pooling layer was also altered to obtain a single number for prediction for each class. Once the saliency maps were stored by the model, WILDCAT pooling was used to average the top *t* percent of pixels in the maps for each class, resulting in a single number used for prediction. WILDCAT pooling is a recently studied alternative to GAP and GMP— discussed in Chapter 2—that balances highlighting only the most important discriminative regions used by a model with the possibility of capturing multiple regions when necessary (Durand et al., 2017).

Lastly, a penalty term was introduced to constrain the saliency maps to only the most important regions. This was done with L_1 regularization following the approach used by Shen, Shamout, et al. (2021). The mean value of each saliency map multiplied by *lambda* was added to the loss function. While Shen, Shamout, et al. tested a range of *lambda* values, hyperparameter tuning was outside the scope of this project. Instead, the model was trained with and without regularization (*lambda = 0.01* and *lambda = 0*, respectively) to see if the maps were improved and more localized to the lesion.

Conformer

The Conformer model used follows the framework described in *TransCAM: Transformer Attention-based CAM Refinement for Weakly Supervised Semantic Segmentation*, a recently published paper proposing TransCAM, a Conformer that segments objects using only image-level labels (Li et al., 2022). TransCAM, which combines a visual transformer branch with a CNN branch, refines the CAM generated from a CNN by leveraging the attention weights

generated by the transformer branch. Because this technique is so new, the model consisted of adapting the freely available TransCAM code provided by the authors to the task of processing grayscale, single-channel breast US images. Pre-trained Conformer weights provided by the original architects of the Conformer were used (Peng et al., 2021), and the data augmentations for training data were altered to match those used for the CNN models.

Training

The CNN and Saliency Models were trained in Google Collab notebooks using the fast.ai library, which allowed for the simple implementation of deep learning networks. Two separate Saliency models were trained, one without regularization (*lambda = 0*), and one with regularization (*lambda = 0.01*). Cross entropy loss was used as the loss function for fine-tuning and fast.ai's function `lr_find()` found the best learning rate for each model. After the learning rates were determined, each model was trained for up to 500 epochs and the model with the highest accuracy was saved. MixUp, a technique to prevent overfitting, was also utilized in the training process. MixUp blends two images at random from the training data—about 50% of pixels from each are represented—and makes it more difficult for the algorithm to memorize particular features of the categories, which prevents overfitting (Zhang et al., 2021).

Following the design of TransCAM, the ViT was trained for 20 epochs with a learning rate of 0.00005, and the loss function soft margin loss. The optimizer called AdamW, which is an extension of stochastic gradient descent, was also used in model training (Kingma and Ba, 2014; Loshchilov & Hutter, 2017). Because the original application was for multilabel classification—that is, detecting instances of multiple classes in one image—the input pathology data was one-hot encoded to generate CAMs and attention maps for each category. The model and parameter weights with the best performance were then saved.

Inference and Visualization

After training and testing the models on the test set consisting of images from all four data sources, only the 55 randomly selected Mayo breast US images in the test set were chosen for subsequent inference and visualization. This was done because the images were of a higher quality and would produce results most relevant to the BUS Project, mimicking the eventual output from the final software.

The CNN model was used to predict pathology labels for each image in the Mayo test set. From there, a CAM object was constructed for each of the most popular methods of pixel explainability. This follows the implementation guide associated with the `pytorch_grad_cam` library used in this research (Gildenblat, 2021). The following methods were applied: (a) GradCAM, weights the two-dimensional activations by the average gradient; (b) GradCAM++ is similar to GradCAM but uses second-order gradients; (c) HiResCAM is similar to GradCAM but performs element-wise multiplication of the activations and gradients, guaranteeing faithfulness; (d) EigenCAM takes the first principle component of the two-dimensional activations, but has no class discrimination; (e) EigenGradCAM is similar to EigenCAM but has class discrimination; (f) FullGrad computes the gradients of the biases from all over the network, and then sums them. For each method, a heatmap was generated showing the discriminative regions activated for the predicted class. This heatmap was then overlaid on the original breast US image, highlighting regions used by the CNN model.

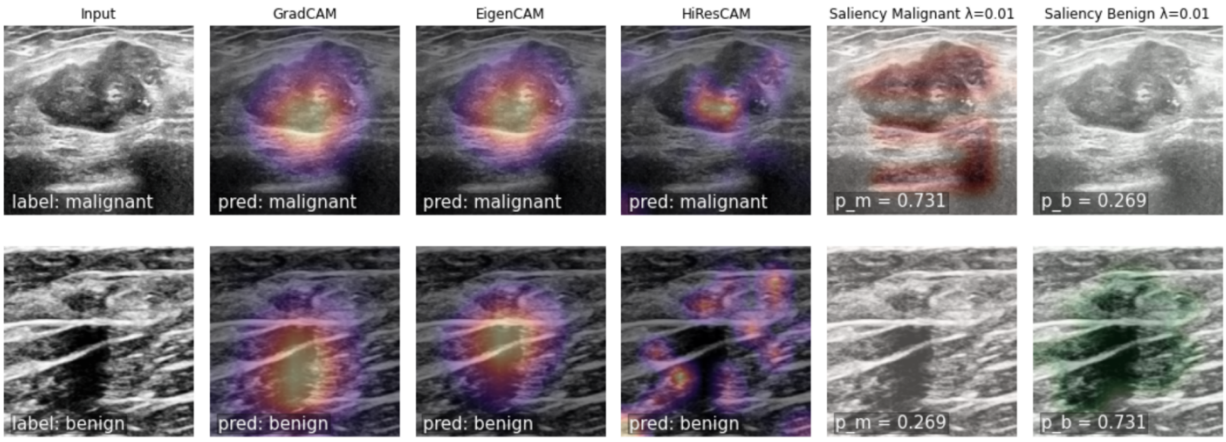
The Saliency model was used to predict pathologies on the same Mayo test set. Because saliency map generation was built into the model architecture, visualizing these maps was a matter of plotting the original image alongside the maps generated for each class. Malignant

activations were shown in red and benign in green. In addition, the probability associated with each class was added to the plot.

Because the procedures for modeling, predicting, and visualizing CAMs and saliency maps were so similar, a combined plot was created to show both methods of model explainability. This can be seen in Figure 8. For simplicity, only a selection of CAM types and Saliency maps were included in the plot.

Figure 8

Visualization of GradCAM, HiResCAM, EigenGradCAM, and Saliency Maps



Note. The CNN and Saliency models predicted the pathology of two Mayo test images containing malignant lesions. The CNN model used for GradCAM, HiResCAM, and EigenGradCAM incorrectly labeled both lesions as benign. The Saliency model with L_1 regularization *lambda = 0.01* correctly identified the lesions as malignant.

Like the saliency map method, attention map-enhanced CAMs were generated when the model was evaluated because they are built into the model architecture. To visualize TransCAM on the Mayo test data, two custom functions provided by Li et al. (2021) were utilized to first extract the heatmap from the model and then plot the heatmap alongside the original image. However, the application of ViTs to breast US is very new and the visualizations generated were

not readily comparable to the previous methods. Consequently, the visualizations from TransCAM were not incorporated into the combined plot above.

Summary

This chapter covered the methods used to collect data, build and train three deep learning algorithms, and create visual explanations for the algorithm's predictions. The limited number of high-quality breast US images provided by Mayo Clinic were supplemented with three public breast US image datasets. Histogram equalization was applied to the images to bring out fine details and the data was split into train and test sets. The features and modifications specific to each model architecture—basic CNN model with ResNet-34 backbone, CNN model modified for saliency maps, and combined CNN and ViT Conformer TransCAM—were described along with the data augmentations employed during model training. Finally, the methods for extracting heatmaps that provide model explainability were demonstrated. A link to the code used in this project can be found in Appendix A.

Chapter 4: Results

Introduction

This chapter reports the most important findings of the study, including the outcomes from each deep learning model and the associated visual explanation method covered in the previous chapter. The quantitative results from model training are discussed in terms of overall accuracy in classifying images as benign or malignant and any variation between classes through confusion matrices. This is followed by a qualitative evaluation of the interpretability methods and the final visualizations produced. Lastly, the most promising method for visually explaining CAD of breast US lesions is summarized.

It should be noted that the BUS Project represents an ongoing effort that could go through many iterations before entering a radiologist's toolkit. The work presented here demonstrates how far the data science community has come in making black box algorithms interpretable and shows that these advancements can be applied to breast US imaging.

Model Performance

A summary of relevant metrics for each model is shown in Table 2. Each model is discussed in more detail below.

Table 2

Comparing Model Performance

Model	Accuracy	False-Positive Rate	Sensitivity	Specificity
Basic CNN	83.4	13.9	79.9	86.1
Saliency Model $\lambda = 0.01$	78.2	20.9	77.1	79.1
TransCAM	73.1	37.4	86.8	62.6

The basic CNN model used to generate CAMs and the modified Saliency Model used to generate saliency maps were built with the pre-trained ResNet-34 backbone and were each trained for up to 500 epochs. The training data contained high-quality breast US images from Mayo Clinic and supplemental images from public datasets, with a very slight class imbalance of more benign lesions than malignant lesions. The models achieving the highest accuracy on the test data—which contained image quantities from each data source proportionate to the training dataset—were saved as the final model.

The best model performance for the basic CNN model was found after 168 epochs of training and achieved 83.4% accuracy in the binary classification task. Table 3 shows the confusion matrix results from this model.

Table 3

Basic CNN Model Confusion Matrix

		Predicted Values	
		benign	malignant
Actual Values	benign	161	26
	malignant	29	115

Table 3 reveals that the model performed slightly worse on malignant lesions, misclassifying more malignant lesions as benign than the reverse. This could be particularly harmful in the medical field as it could mean some cancers go undetected. The false-positive rate, or rate of benign images classified as malignant, was only 13.9%.

The custom Saliency model altered the basic CNN model by generating saliency maps for each class and then applying WILDCAT pooling instead of GAP to produce a single value for prediction from each map. Two Saliency models were trained, one with L_1 regularization (*lambda = 0.01*) that added a penalty term to the cross-entropy loss function to constrain the saliency maps to the most important regions, and one without regularization (*lambda = 0*). While both models achieved 78.2% accuracy, results are only shown for the model with regularization *lambda = 0.01* as it had a lower false-positive rate. Table 4 shows the confusion matrix results from this model. The confusion matrix for the model without regularization can be found in Appendix B.

Table 4

Saliency Model with L_1 Regularization Confusion Matrix

		Predicted Values	
		benign	malignant
Actual Values	benign	148	39
	malignant	33	111

The Saliency model performed slightly worse than the basic CNN model, most likely due to the sigmoid function and altered pooling method. While the CAM visualization method is applied retroactively without altering model architecture, the saliency map affects the classification results. The false-positive rate was 20.9%, seven points higher than the basic CNN, and the Saliency model had the opposite difficulty regarding the classes, misclassifying more benign lesions as malignant than the reverse. However, this model misclassified more benign *and* malignant lesions compared to the basic CNN model, which could be weighed against its viability for model visualization.

The final deep learning algorithm, TransCAM, was trained according to the paper by Li et al. (2021) using pre-trained weights from the creators of the Conformer (Peng et al., 2021). After 20 epochs of training using the AdamW optimizer, the model was saved. It was then tested on the test set of images and achieved an accuracy of 73.1%. Worth noting, the inference framework described by Li et al. was geared toward pixel-level classification rather than image-level classification. To predict and validate image-level labels, the pseudo-labels generated for

optimizing the loss function—found by combining the logits from the transformer and CNN branches—were used to calculate accuracy. Table 5 shows the confusion matrix from this model.

Table 5

TransCAM Confusion Matrix

		Predicted Values	
		benign	malignant
Actual Values	benign	117	70
	malignant	19	125

As can be seen, the TransCAM model achieved lower accuracy than the other models on the test data and had a false-positive rate of 37.4%, sixteen points higher than the Saliency model. While the model correctly classified more malignant lesions than the previous two models, it significantly over-predicted benign images as malignant, leading to a high false-positive rate. These difficulties could be due to the fact that image-level classification was not included in the original model framework and was instead derived simplistically from the available code. In addition, the original multi-label classification loss function was changed to cross-entropy loss for this project and the Dense CRF used to improve pseudo-labels was not employed. Finally, no method was used to repeatedly fit the model and save the model with the highest accuracy. After training, the model was simply run on the test data, and the results were recorded.

The objective for building and training a deep learning algorithm to classify lesions in breast US images was to achieve 80-90% accuracy on test data using a deep learning model

architecture. This level would be considered state-of-the-art (Cao et al., 2019) and would match board-certified radiologist performance. While the basic CNN model achieved this goal, the Saliency model and TransCAM fell slightly short. Given that only basic data augmentations and simplified model optimization techniques were used to train the models, the accuracies achieved are acceptable in the context of this study for simulating the CAD software and testing visual explanation methods.

Visualization Methods

After achieving the accuracies shown in Table 2, each model was used for inference on the 54 Mayo images included in the test set. This was done to obtain predicted classification labels and heatmaps for each image, mimicking the real-world implementation of BUS Project CAD software.

Class Activation Maps

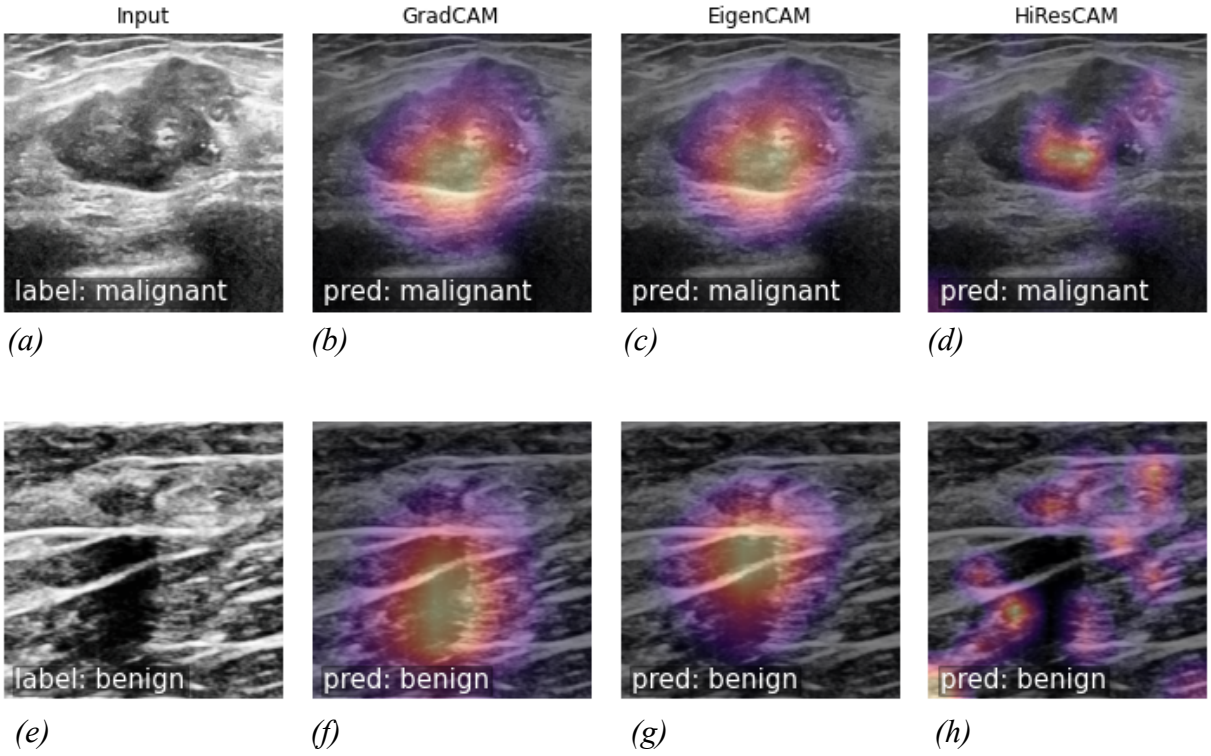
The first visualization method, class activation map (CAM), was implemented using the basic CNN model architecture and applied many of the popular AI explainability methods contained in the `pytorch_grad_cam` library (Gildenblat, 2021). Of the six methods tested on Mayo images—GradCAM, GradCAM++, HiResCAM, EigenCAM, EigenGradCAM, and FullGrad—only three produced distinct, meaningful heatmaps highlighting regions used for classification. These three methods can be seen in Figure 9 on a sample of Mayo test images. GradCAM and EigenCAM highlighted very similar regions of the image, whereas HiResCAM created a more patchwork-like heatmap.

GradCAM and EigenCAM both appear to locate and highlight the lesions contained in each breast US image. These methods should give a radiologist confidence in the model's ability to find the region of interest and use it for classification. Image (f) shows that GradCAM

captured the elongated shape of the lesion, while EigenCAM (g) did not. And while the margins of the lesion are not precisely outlined, Dr. Rich Ellis, the Mayo Clinic radiologist attached to the BUS Project, noted that oftentimes the borders of a lesion are not well-defined, and so visualizing a precise boundary is not always necessary. Additional examples from all six methods on more Mayo images can be found in Appendix C.

Figure 9

GradCAM, EigenCAM, and HiResCAM on Mayo Breast US Images



Note. Because all CAM methods are applied after the CNN model classified the image, all methods show the same predicted class.

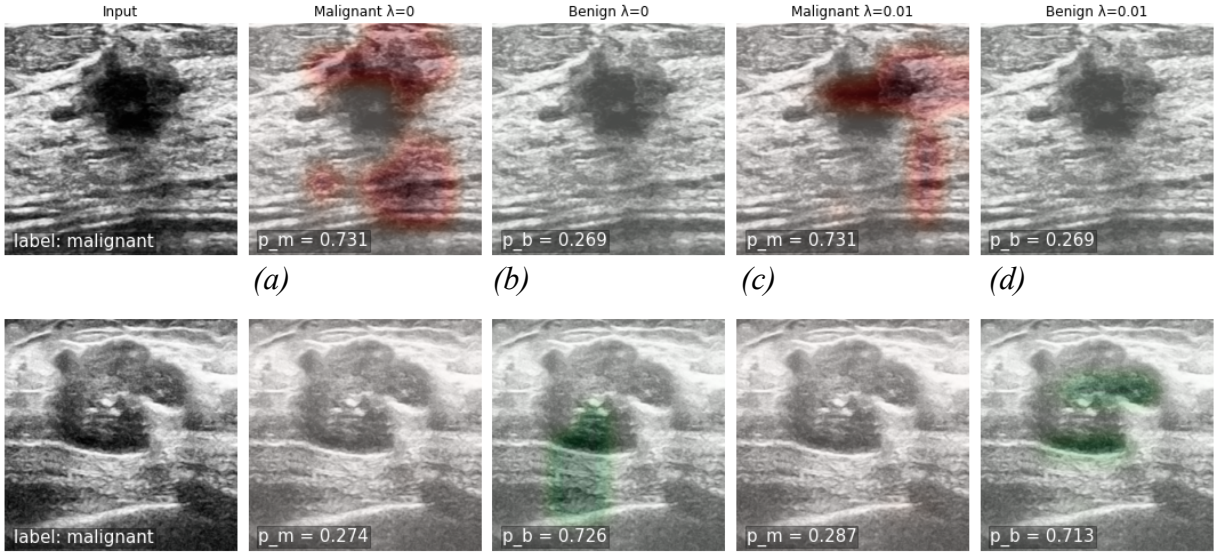
Saliency Maps

The second visualization method, saliency maps, was implemented through the Saliency model architecture. Two saliency maps, one for each class, with values representing the

contribution of each pixel toward the prediction of a malignant or benign lesion, were generated for each Mayo test image. The maps were then plotted as heatmaps over the original Mayo images. Figure 10 displays a sample of these visualizations comparing the models trained with and without L_1 regularization.

Figure 10

Saliency Maps on Mayo Breast US Images



Note. The labels at the bottom of the saliency maps show the probabilities calculated by the models. Heatmaps could appear on both the Malignant and Benign maps if the probabilities were closer to 0.50. The columns with $\lambda = 0$ are from the model trained without regularization.

The models correctly classified both images and highlighted portions of each lesion. Regions with greater color intensity, red or green, showed where the model found the most important information for classification. The benign saliency map in (h) captured most of the lesion and showed increased intensity in the center of the lesion. This image should give a radiologist confidence in the model's detection and classification of the lesion. However the malignant saliency map in (c) highlighted some regions outside the lesion, particularly in the far-

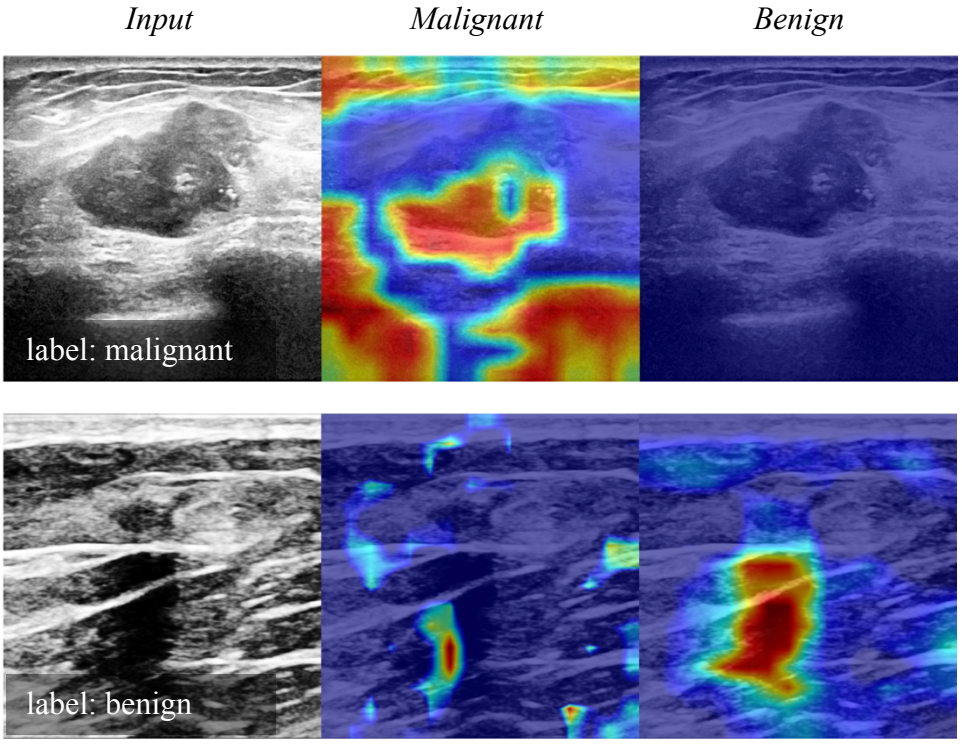
right side of the image. This would be useful information for the radiologist to have: even though a large portion of the lesion was used by the model to predict malignancy, extraneous pixels were used as well. Overall, the saliency maps with *lambda = 0.01* were more constrained to the lesions, highlighting fewer outside regions than the maps without regularization. From here, the saliency map method will refer only to the model trained with L_1 regularization.

Attention Map-Enhanced Class Activation Maps

The final visualization method, attention-map enhanced CAMs, was implemented with TransCAM. The model architecture produced both classification labels and CAMs that were refined with the attention maps generated by the transformer branch of the model. Two maps were produced for each Mayo test image, one with the malignant activations and one with the benign activations. Figure 11 displays a sample of the visualizations produced.

Figure 11

TransCAM Attention Enhanced CAM on Mayo Breast US Images



TransCAM highlighted the lesions from both images, but also highlighted unrelated artifacts and shadows that were not parts of the lesion. These images, as they are, would not give a radiologist confidence in the model's classification. However, TransCAM does appear to outperform both the basic CNN and Saliency models in capturing the borders and irregular shapes of the lesions. While this type of model still needs refinement and more customization to the task of breast US, these results show the promising possibility of more precise lesion localization or segmentation.

Conclusion

The three model architectures achieved acceptable levels of accuracy for the purpose of this study. While it was possible to produce meaningful visualizations of the regions used by each model for classification, those visualizations could improve with higher-performing models. Of the three visualization techniques, saliency maps provided the best combination of model performance and interpretability. Each method had strengths and weaknesses, but the Saliency model generated the most meaningful heatmaps that communicated the most class-specific information. Moreover, the model has the possibility of improved performance when more data becomes available. The CAMs often produced heatmaps that highlighted very large and non-specific regions, and the attention-map enhanced CAMs from TransCAM were not yet of high enough quality or confidence to match the other two methods.

Chapter 5: Discussion

Introduction

The purpose of this study was to research, implement, and evaluate some of the most popular methods for visually explaining deep learning algorithms in application to breast US image classification. The American College of Radiology advocates for radiologists to have a

high-level understanding of an AI model when used in a clinical setting (Sendak et al., 2021). One way to address this is by visualizing the regions of the US used in classification. This chapter presents a summary and discussion of the results, including limitations and the potential consequences, as well as recommendations for future research.

Summary of Findings

Three deep learning algorithms were trained and tested on 1678 breast US images from the Mayo Clinic and from public datasets. For each algorithm, a method of model visualization was then implemented to interpret how the model predicted lesion pathology. These visualizations all took the form of heatmaps highlighting regions of the image used by the algorithm. Saliency maps with L1 regularization, created by the Saliency model, were selected as the most promising visual tool for a radiologist to understand and interpret the so-called black box deep learning algorithm. The model achieved 78.2% accuracy on the test set and the heatmaps were shown to localize the lesions and differentiate between classes. The class-specific information communicated by the saliency maps set them apart from the single-image CAMs—and while the attention map-enhanced CAMs had the potential to communicate similar class-specific information, they were not fully developed in this study and the quality was not at a level to compare to the other methods. In addition, because this study focused on how information could be presented to a radiologist, the <80% accuracy was acceptable within the scope of the project.

Discussion

While the accuracy levels achieved by the models in this project were slightly below the objective of 80-90%, the models performed well enough to show what is possible for visually explaining the deep learning algorithms. Of the three models, the basic CNN with ResNet-34

backbone is most similar to the architecture currently favored by the BUS Project, albeit simplified to predict a binary pathology rather than a BI-RADS score. Additionally, this model reached the highest accuracy at 83.4%. However, the popular CAM techniques that can be applied to any CNN without altering the model architecture produced simple visualizations that did little more than locate the area of the lesion. For a radiologist, these visual explanations would only demonstrate whether or not the model found the lesion.

Alternatively, ViTs are gaining more attention within the BUS Project and in the CAD field more broadly (Matsoukas et al., 2021). The implementation of TransCAM shown in this project was exploratory in nature, and has much room for improvement. That being said, the attention map-enhanced CAMs marked the borders of the lesions with much more detail than both the CAMs and saliency maps. With further development, these visualizations could be a valuable model explainability tool if the BUS Project moves in the direction of replacing CNNs with ViTs.

Overall, the finding that saliency maps are the most promising visual tool for explaining the CNN model supports the studies by Shen, Shamout, et al. (2021) and Shen, Wu, et al. (2021). These two studies using AI to identify breast cancer in US and mammography utilized saliency maps both to localize the ROI and to explain the predictions. Given only image-level labels, the models were able to locate malignant lesions. This develops radiologist trust and helps them understand the model's strengths and limitations. The saliency maps produced by this study show the same possibility for the BUS Project.

One main challenge this project faced was the lack of data. With more high-quality breast US images from Mayo clinic, the models could have achieved higher accuracies and produced heatmaps that more specifically located the lesions. For example, many of the saliency maps did

not clearly define the borders of the lesions. With more data and more rigorous optimization, the Saliency model should be able to achieve accuracy on par with other state-of-the-art deep learning CAD systems.

Another important consideration is the high false-positive rate of the Saliency model. Compared to the study mentioned above, which focused on AI reducing false-positive diagnoses, the Saliency model underperformed (Shen, Shamout, et al., 2021). This is a weakness of the model that will need to be closely studied and accounted for if the BUS Project moves forward in implementing a similar model architecture. It is possible that this is a result of the simplified model implementation coupled with the lack of data, but it could also represent a major shortcoming of the method.

Suggestions for Future Research

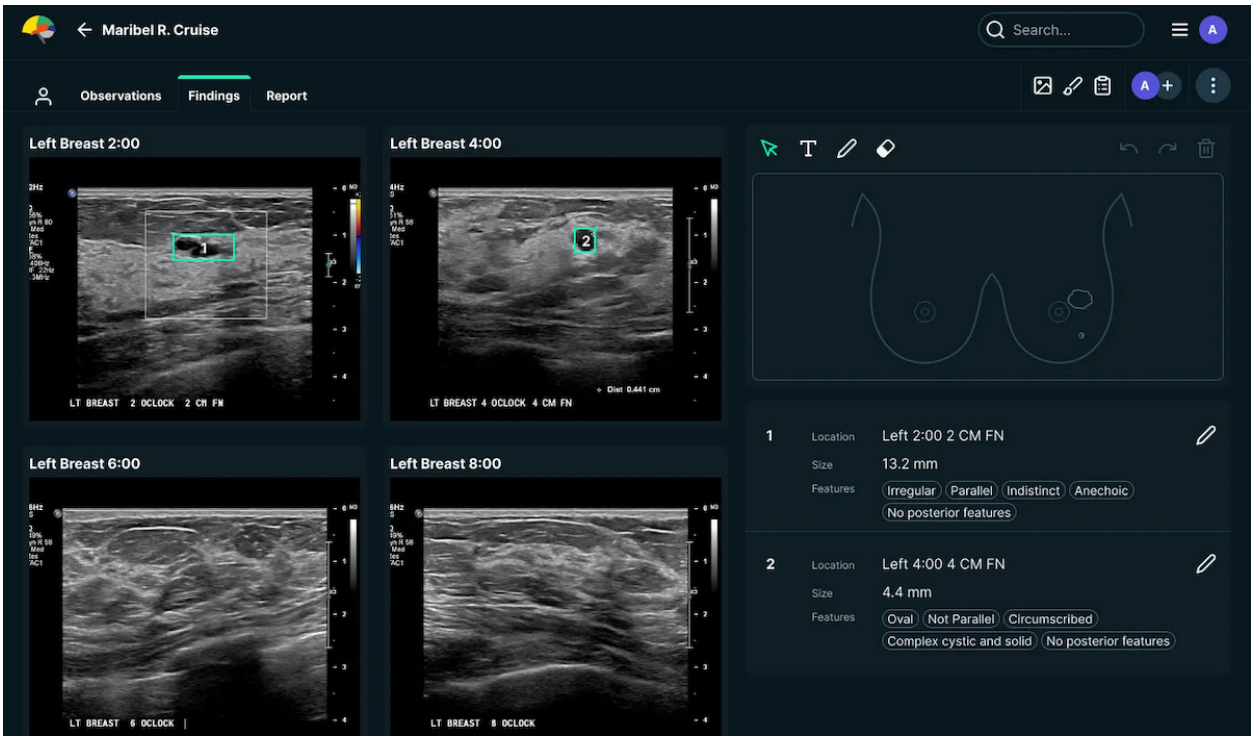
The viability of CAMs as a simple visual tool and saliency maps as a more informative tool for explainable AI applied to breast US have been demonstrated in this study. In addition, attention map-enhanced CAMs were shown to have a promising ability to highlight the irregular borders of lesions, although the overall model achieved lower accuracy and often highlighted extraneous regions of the image. Due to time and data limitations, further refinement of these methods was not within the scope of this project but is recommended for future research.

First, the method for constraining the saliency maps to highlight only the most important pixels should be tested and developed further. The basic comparison with and without L_1 regularization was only the first iteration of this goal. In their study, Shen, Shamout, et al. (2021) tested $lambda \in 10^{[-3,\, 0.5]}$ for hyperparameter tuning to create the best saliency maps. In a different approach, the Med-Tech startup See-Mode recently released an AI system to help radiologists find and classify lesions according to the BI-RADS score, and its software appears

to use bounding boxes to help localize the lesions (Ang, 2022). Figure 12 shows an example of their system.

Figure 12

See-Mode Radiologist Interface (Ang, 2022)

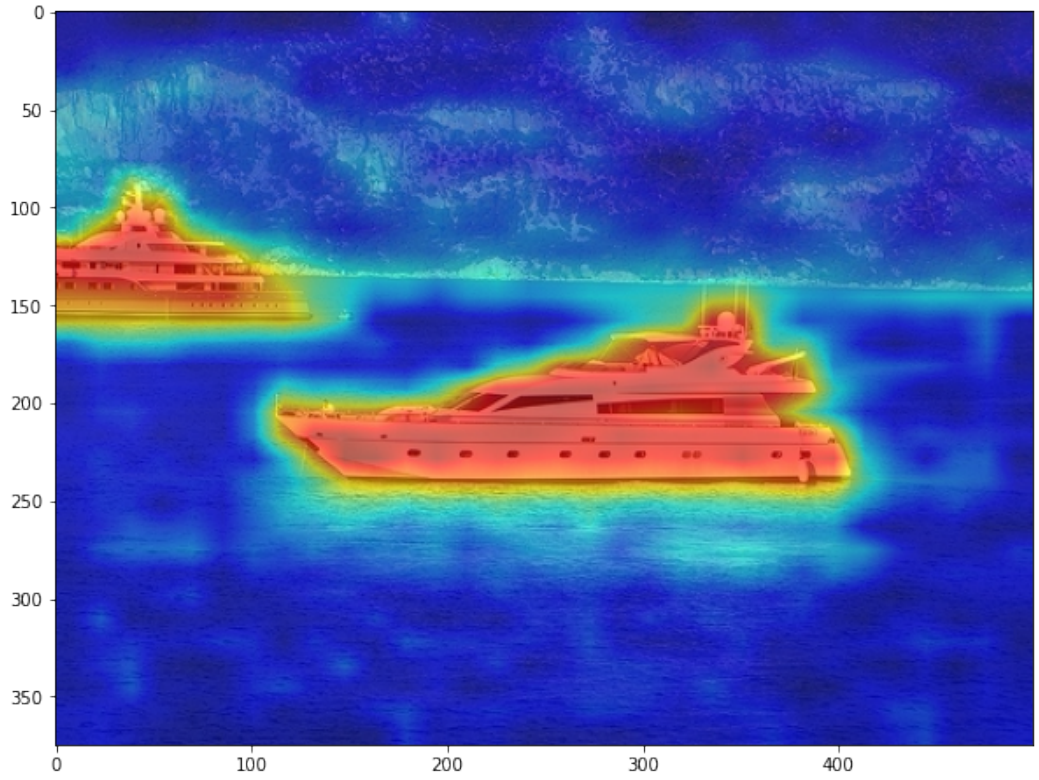


While model explainability does not appear to be a major feature on the interface, precise bounding boxes contain the lesions and exclude extraneous regions of the US image. If combined, these two techniques could improve the readability of saliency maps for radiologists and better localize the lesions.

Second, further research in applying TransCAM to breast US is necessary if the BUS Project moves in the direction of ViTs. The potential for attention map-enhanced CAMs goes beyond what has been seen in any of the other visualization methods for segmenting objects and following the borders. Figure 13 demonstrates what the model achieved on a generic image dataset when asked to highlight the “boat” class.

Figure 13

TransCAM Detecting "Boat" (Li et al., 2022)



Not only are the boats almost perfectly outlined, but two are highlighted in one image. This stems from the multilabel setup of the data used for TransCAM. While the breast US data was set up in the same way for this study, none of the images actually contained more than one lesion. However, in a clinical setting, there may be more than one lesion present in a US image. Additionally, the Dense CRF used in the original study to refine pseudo-labels was not used in this project. The study and implementation of the many optimization techniques used in TransCAM could result in compelling visualizations that push the BUS Project in the direction of ViTs.

Finally, an additional study is necessary to fully implement one or multiple of the studied visualization techniques on the full model architecture used by BUS Project. The input will

include multiple characteristics of the lesion rather than just a pathology label, and the model output will be a BI-RADS score. The more complex model could affect the clarity and confidence of the visualization methods that work well within the scope of binary pathology labels.

Conclusion

This study aimed to implement and evaluate three methods for visually explaining deep learning algorithms when applied to breast US classification. The results showed that saliency maps produced by a modified CNN model architecture communicate the most information. The class-specific heatmaps would provide a radiologist with confidence in the black box algorithm when the lesion is localized, and reveal limitations in the model prediction when extraneous regions are highlighted. With additional data and improved model optimization techniques, this model could achieve state-of-the-art accuracy and produce more accurate and precise visualizations than those shown in this study. As more data becomes available and the CAD system is improved, it is the hope of this study that the BUS Project continues to develop and prioritize model explainability for the sake of responsible AI in healthcare.

References

Agarwal, S. (2022, June 3). More than Just AI: Practical Applications in Diagnostic Intelligence. *HealthTech*. https://healthtechmagazine.net/article/2022/06/more-just-ai-practical-applications-diagnostic-intelligence

Ang, A. (2022, November 16). Medtech startup See-Mode gets Canada's nod for AI breast, thyroid ultrasound analysis software. *MobiHealthNews*. https://www.mobihealthnews.com/news/asia/medtech-startup-see-mode-gets-canada-s-nod-ai-breast-thyroid-ultrasound-analysis-software

Al-Dhabyani, W., Gomaa, M., Khaled, H., & Fahmy, A. (2020). Dataset of breast ultrasound images. *Data in Brief, 28*, 104863–104863. https://doi.org/10.1016/j.dib.2019.104863

Berg, W. A., Bandos, A. I., Mendelson, E. B., Lehrer, D., Jong, R. A., & Pisano, E. D. (2015). Ultrasound as the Primary Screening Test for Breast Cancer: Analysis From ACRIN 6666. *Journal of the National Cancer Institute*, *108*(4), djv367. https://doi.org/10.1093/jnci/djv367

Cao, Z., Duan, L., Yang, G., Yue, T., & Chen, Q. (2019). An experimental study on breast lesion detection and classification from ultrasound images using deep learning architectures. *BMC medical imaging*, *19*(1), 51. https://doi.org/10.1186/s12880-019-0349-x

Chefer, H., Gur, S., & Wolf, L. (2020). *Transformer Interpretability Beyond Attention Visualization*. ArXiv preprint. https://doi.org/10.48550/arXiv.2012.09838

Cole, P. (2020, October 16). What's Under the Hood of Neural Networks? *InsideBIGDATA*. https://insidebigdata.com/2020/10/16/whats-under-the-hood-of-neural-networks/

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *International Conference on Learning Representations.* https://doi.org/10.48550/arXiv.2010.11929

Draelos, R. L., & Carin, L. (2020). *Use HiResCAM instead of Grad-CAM for faithful explanations of convolutional neural networks.* ArXiv preprint. https://doi.org/10.48550/arXiv.2011.08891

Evans, A., Trimboli, R. M., Athanasiou, A., Balleyguier, C., Baltzer, P. A., Bick, U., Camps Herrero, J., Clauser, P., Colin, C., Cornford, E., Fallenberg, E. M., Fuchsjaeger, M. H., Gilbert, F. J., Helbich, T. H., Kinkel, K., Heywang-Köbrunner, S. H., Kuhl, C. K., Mann, R. M., Martincich, L., & Panizza, P. (2018). Breast ultrasound: recommendations for information to women and referring physicians by the European Society of Breast Imaging. *Insights into imaging*, *9*(4), 449–461. https://doi.org/10.1007/s13244-018-0636-z

Giaquinto, A. N., Sung, H., Miller, K. D., Kramer, J. L., Newman, L. A., Minihan, A., Jemal, A., & Siegel, R. L. (2022). Breast Cancer Statistics, 2022. *CA: A Cancer Journal for Clinicians*. https://doi.org/10.3322/caac.21754

Gildenblat, J. (2021). *PyTorch library for CAM methods*. GitHub. https://github.com/jacobgil/pytorch-grad-cam

Ginsburg, O., Yip, C. -H., Brooks, A., Cabanes, A., Caleffi, M., Dunstan Yataco, J. A., Gyawali, B., McCormack, V., McLaughlin de Anderson, M., Mehrotra, R., Mohar, A., Murillo, R., Pace, L. E., Paskett, E. D., Romanoff, A., Rositch, A. F., Scheel, J. R., Schneidman, M.,

Unger-Saldaña, K., . . . Anderson, B. O. (2020). Breast cancer early detection: A phased approach to implementation. *Cancer*, *126*(S10), 2379-2393. https://doi.org/10.1002/cncr.32887

IBM Cloud Education. (2020, October 20). *Convolutional Neural Networks*. IBM. https://www.ibm.com/cloud/learn/convolutional-neural-networks

International Agency for Research on Cancer. (2021, February 4). World Cancer Day: Breast cancer overtakes lung cancer in terms of number of new cancer cases worldwide. IARC showcases key research projects to address breast cancer [Press release]. https://www.iarc.who.int/wp-content/uploads/2021/02/pr294_E.pdf

Jarvey, J. L. (2022). *Computer-Aided Detection (CAD) System for Breast Ultrasound Lesion Interpretation: An Explainable Deep Learning Approach* [Unpublished doctoral dissertation/master's thesis]. University of Wisconsin - La Crosse.

Jiménez-Gaona, Y., Rodríguez-Álvarez, M. J., & Lakshminarayanan, V. (2020). Deep-Learning-Based Computer-Aided Systems for Breast Cancer Imaging: A Critical Review. *Applied Sciences*, *10*(22). https://doi.org/10.3390/app10228298

Kim, J., Kim, H. J., Kim, C., & Kim, W. H. (2021). Artificial intelligence in breast ultrasonography. *Ultrasonography (Seoul, Korea)*, *40*(2), 183–190. https://doi.org/10.14366/usg.20117

Kingma, D. P., & Ba, J. (2014). *Adam: A Method for Stochastic Optimization.* ArXiv preprint. https://doi.org/10.48550/arxiv.1412.6980

Li, R., Mai, Z., Trabelsi, C., Zhang, Z., Jang, J., & Sanner, S. (2022). *TransCAM: Transformer Attention-based CAM Refinement for Weakly Supervised Semantic Segmentation.* https://doi.org/10.48550/arxiv.2203.07239

Lin, Z., Lin, J., Zhu, L., Fu, H., Qin, J., & Wang, L. (2022). A New Dataset and a Baseline Model for Breast Lesion Detection in Ultrasound Videos. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022* (pp. 614–623). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-16437-8_59

Loshchilov, I., & Hutter, F. (2017). Decoupled Weight Decay Regularization. https://doi.org/10.48550/arxiv.1711.05101

Matsoukas, C., Haslum, J. F., Söderberg, M., & Smith, K. (2021). Is it Time to Replace CNNs with Transformers for Medical Images? https://doi.org/10.48550/arXiv.2108.09038

Mendelson, E. B., Bôhm-Vélez, M., Berg, W. A., Whitman, G. J., Feldman, M. I., Madjar, H., Rizzatto, G., Baker, J. A., Zuley, M., Stavros, A. T., Comstock, C., & van Duyn Wear, V. (2013). Ultrasound. In *ACR BI-RADS Atlas, Breast Imaging Reporting and Data System* (5th ed., p. 123). American College of Radiology.

NVIDIA. (n.d.). *Convolutional Neural Network*. https://www.nvidia.com/en-us/glossary/data-science/convolutional-neural-network/

OpenCV. (n.d.). *Histograms - 2: Histogram Equalization*. https://docs.opencv.org/4.x/d5/daf/tutorial_py_histogram_equalization.html

Peng, Z., Huang, W., Gu, S., Xie, L., Wang, Y., Jiao, J., & Ye, Q. (2021). *Conformer: Local Features Coupling Global Representations for Visual Recognition.* ArXiv preprint. https://doi.org/10.48550/arxiv.2105.03889

Qi, X., Zhang, L., Chen, Y., Pi, Y., Chen, Y., Lv, Q., & Yi, Z. (2019). Automated diagnosis of breast ultrasonography images using deep neural networks. *Medical Image Analysis, 52*, 185–198. https://doi.org/10.1016/j.media.2018.12.006

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2020). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *International Journal of Computer Vision, 128*(2), 336–359. https://doi.org/10.1007/s11263-019-01228-7

Sendak, M. P., Roth, C. J., & Balu, S. (2021, December 19). *AI Transparency*. American College of Radiology. https://www.acr.org/Practice-Management-Quality-Informatics/ACR-Bulletin/Articles/January-2022/AI-Transparency

Shen, Y., Shamout, F. E., Oliver, J. R., Witowski, J., Kannan, K., Park, J., Wu, N., Huddleston, C., Wolfson, S., Millet, A., Ehrenpreis, R., Awal, D., Tyma, C., Samreen, N., Gao, Y., Chhor, C., Gandhi, S., Lee, C., Kumari-Subaiya, S., . . . Geras, K. J. (2021). Artificial Intelligence System reduces false-positive findings in the interpretation of breast ultrasound exams. *Nature Communications*, *12*(1). https://doi.org/10.1038/s41467-021-26023-2

Shen, Y., Wu, N., Phang, J., Park, J., Liu, K., Tyagi, S., Heacock, L., Kim, S. G., Moy, L., Cho, K., & Geras, K. J. (2021). An interpretable classifier for high-resolution breast cancer screening images utilizing weakly supervised localization. *Medical Image Analysis, 68,* 101908–101908. https://doi.org/10.1016/j.media.2020.101908

Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). *Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps*. https://doi.org/10.48550/arXiv.1312.6034

Yap, M. H., Pons, G., Marti, J., Ganau, S., Sentis, M., Zwiggelaar, R., Davison, A. K., & Marti, R. (2018). Automated Breast Ultrasound Lesions Detection Using Convolutional Neural Networks. *IEEE Journal of Biomedical and Health Informatics, 22*(4), 1218–1226. https://doi.org/10.1109/JBHI.2017.2731873

Yu, H., Yang, L. T., Zhang, Q., Armstrong, D., & Deen, M. J. (2021, July 15). Convolutional neural networks for medical image analysis: State-of-the-art, comparisons, improvement and perspectives. *Neurocomputing*, *444*, 92-110. https://doi.org/10.1016/j.neucom.2020.04.157

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2015). Object Detectors Emerge in Deep Scene CNNs. *International Conference on Learning Representations*. https://doi.org/10.48550/arXiv.1412.6856

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2016). Learning Deep Features for Discriminative Localization. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2921-2929. https://doi.org/10.1109/CVPR.2016.319

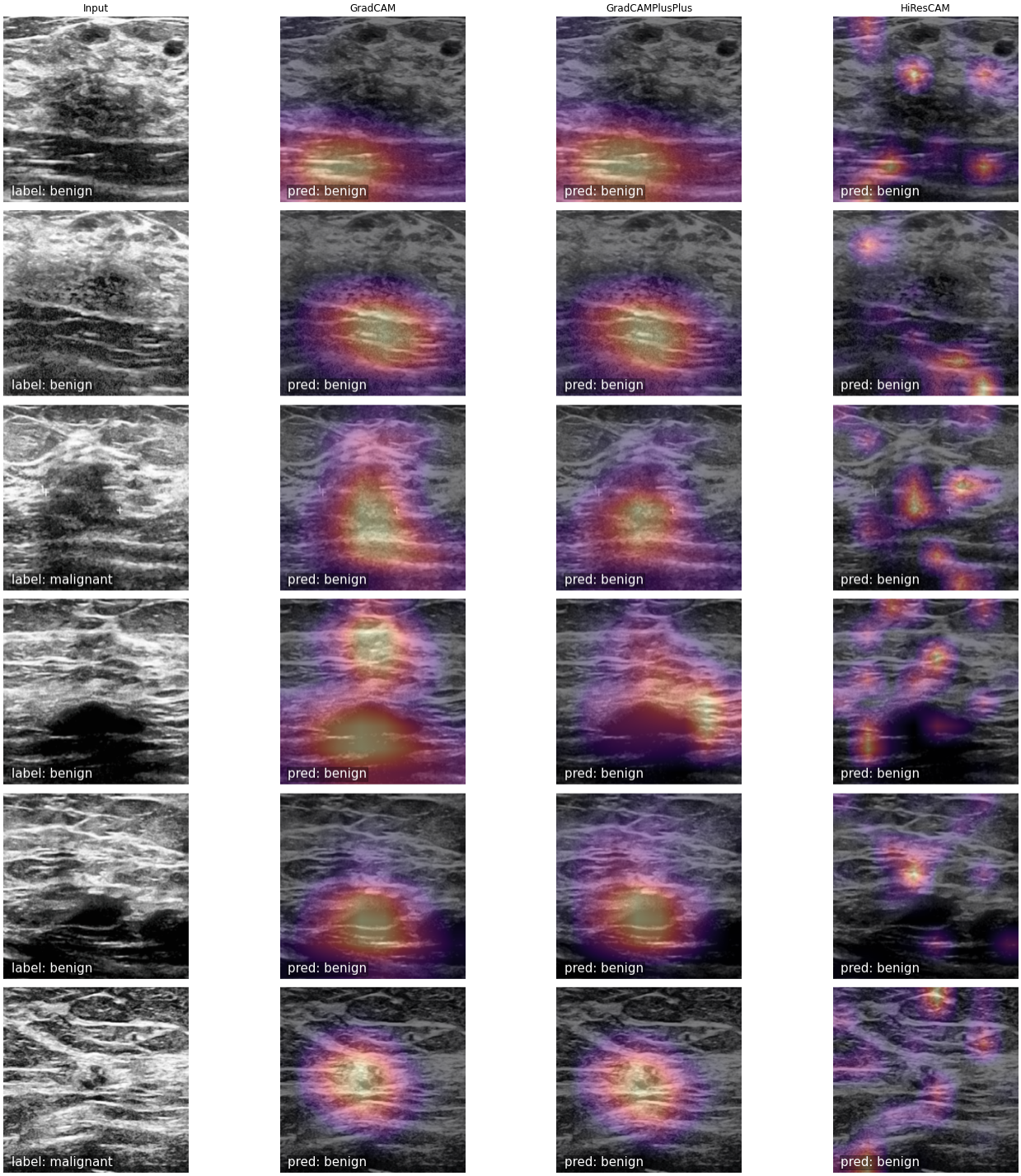
Appendix A: Code

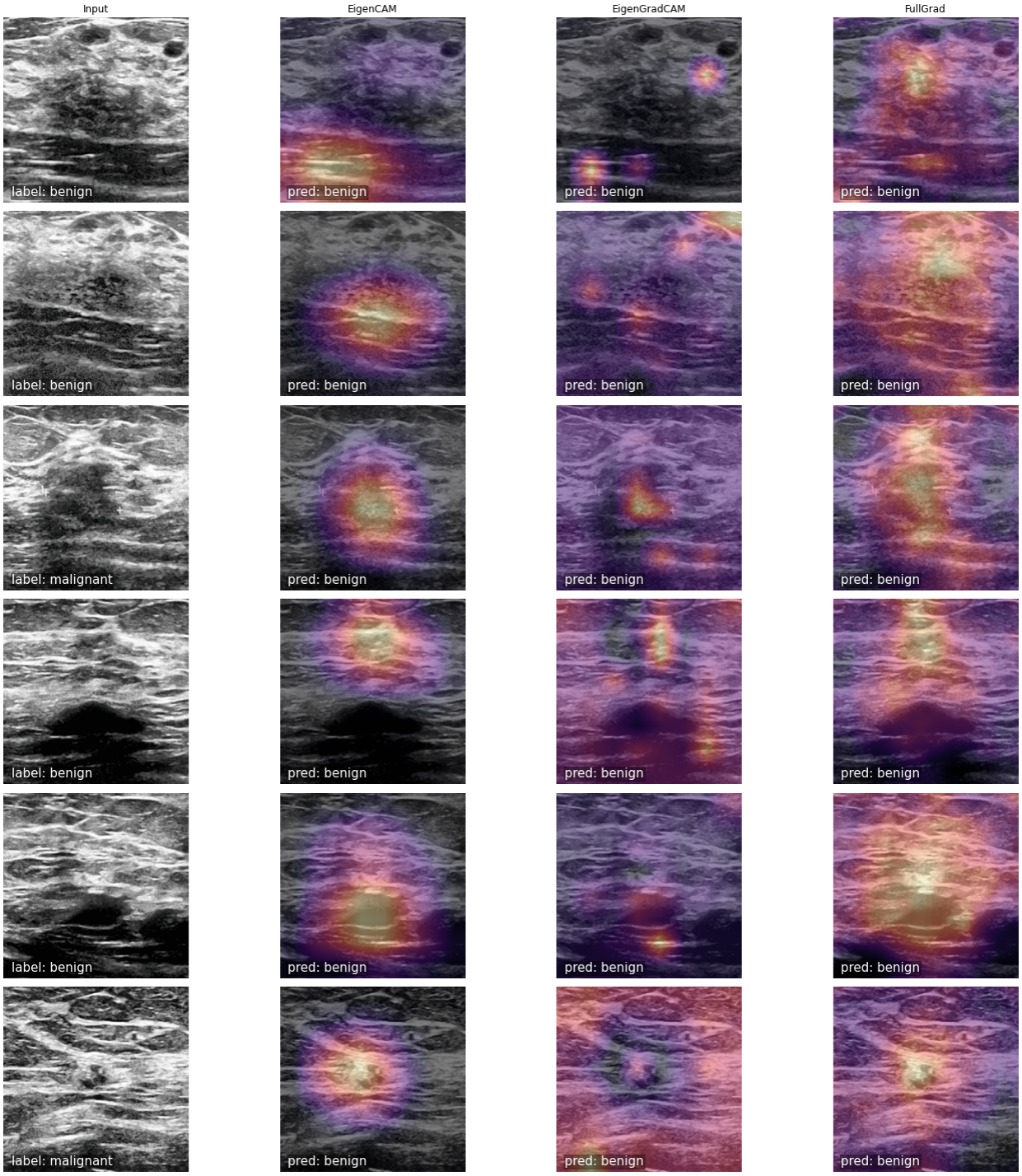
https://github.com/tbodart/capstone/blob/main/Teresa_Bodart_Capstone_Final.ipynb

Appendix B: Saliency Model without Regularization Confusion Matrix

		Predicted Values	
		benign	malignant
Actual Values	benign	146	41
	malignant	31	113

Appendix C: Additional CAM Methods




Input
EigenCAM
EigenGradCAM
FullGrad
label: benign
pred: benign
pred: benign
pred: benign
label: benign
pred: benign
pred: benign
pred: benign
label: malignant
pred: benign
pred: benign
pred: benign
label: benign
pred: benign
pred: benign
pred: benign
label: benign
pred: benign
pred: benign
pred: benign
label: malignant
pred: benign
pred: benign
pred: benign