Computer-Aided Detection (CAD) System for Breast Ultrasound Lesion Interpretation: An Explainable Deep Learning Approach

By

Josh L. Jarvey

A Capstone Project Paper Submitted in Partial Fulfillment of the

Requirements for the Degree of

Master of Science

In

Data Science

University of Wisconsin - La Crosse

La Crosse, Wisconsin

August 2022

Acknowledgements

First, I'd like to thank Dr. Jeff Baggett from the University of Wisconsin – La Crosse for all his guidance throughout the process of this project. His ability to translate data science concepts to code and teach me about its implementation was invaluable and inspired me to try and learn new techniques. I'd also like to thank Dr. Song Chen from the University of Wisconsin – La Crosse for sharing his knowledge and providing his expert recommendations each week during our meetings.

Secondly, I'd like to thank the entire Mayo Clinic Health System for funding this project under the inaugural Collaborative Seed Grant Program. Specifically, I want to thank Dr. Rich Ellis of Mayo Clinic for providing his expertise in radiology and breast imaging. His efforts in gathering medical images, annotating, and classifying the data are what made this project possible.

Next, I'd like to thank the other student researchers on this project: Simon Wagner, Eric Jahns, and Justin Hall. Not only were our interactions amicable, but they also helped me expand my knowledge of software development and deep learning.

Finally, I'd like to thank my wife Carlie, our two daughters Kennedy and Kinley, and my mother. Your constant support and interest in my research gave me the motivation to get through my studies. Lastly, to my late father - I know you'd be proud to hear about the important contributions that were made to the greater community in light of this effort.

Abstract

According to the World Health Organization, breast cancer is now the most prevalent type of cancer diagnosed around the world. In 2020, it accounted for the deaths of roughly 685,000 women alone. Breast cancer incidence and mortality rates are expected to continue to rise in the coming decades, primarily in low to middle-income counties, due in part to an adoption of a more western lifestyle coupled with misconceptions about the nature and curability of the disease. Therefore, in partnership between the University of Wisconsin – La Crosse and Mayo Clinic Health Systems, the purpose of this study is to assess the feasibility of utilizing deep learning to aid radiologists with the interpretation of lesions discovered in breast ultrasound (BUS) images during routine clinical screenings. Based on a review of the literature on medical imaging and computer-assisted detection (CAD) systems for BUS interpretation, a multitask learning model using a pre-trained state-of-the-art convolutional neural network (CNN) was developed and trained using various image augmentation techniques known to increase performance. The research found that the best model identified from this study performed on par with that of a trained radiologist in its ability to predict lesion pathology. However, no definitive conclusions could be drawn about the model's multitask performance due in part to the limited data available. Further research is needed as more data is made available from the Mayo Clinic, and alternative explainability methods may need to be explored.

Keywords: Breast cancer, deep learning, computer vision, multitask learning, explainability, ultrasound, medical imaging, radiology

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Background

According to the World Health Organization (2021), breast cancer is now the most prevalent type of cancer diagnosed around the world. In 2020, it accounted for roughly the deaths of 685,000 women alone. In the United States, breast cancer is the second most commonly diagnosed female cancer, and in 2022, the American Cancer Society (2022) has projected that about 43,250 women are likely to pass away due to the disease. According to DeSantis et al. (2015), breast cancer incidence and mortality rates will continue to increase in the coming decades, primarily in low to middle-income counties, due in part to an adoption of a more western lifestyle coupled with misconceptions about the nature and curability of the disease.

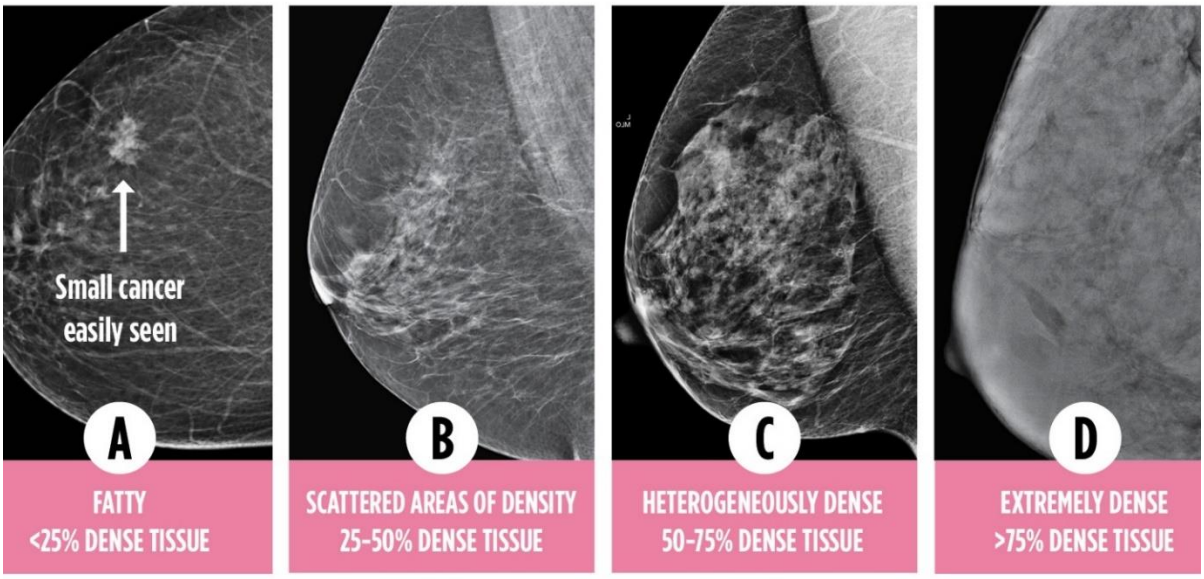
While breast cancer continues to be one of the greatest challenges of our time, there are well-known factors that increase the survivability of the disease. Early detection is chief among these factors known to be directly correlated with higher rates of survivability (Berry et al., 2005; Birnbaum et al., 2018). Although some methods of early detection focus on the diagnosis and downstaging of symptomatic cancer, this project emphasizes the detection of asymptomatic cancer via clinical screening techniques.

A mammogram is a common screening technique that uses low-dose radiation in an attempt to highlight regions of interest within the breast and has shown high levels of efficacy in the early detection of asymptomatic breast cancer (Berry et al., 2005). However, not all patients have access to the technology, and for some, mammography is not an option. One such reason is based on the physiological makeup of the breast tissue. According to Joy et al. (2005), women with radiographically dense breasts, i.e., composed of more connective and epithelial tissue vs. the less dense fatty tissue, are more at risk of a false-negative outcome when undergoing a mammogram screening. This is due to the

similarities of how the radiologically dense tissue compares to calcifications and masses in the imaging results. Figure 1 below compares the results of a mammogram under different classifications of tissue densities. It is much easier to see the cancerous mass in the fattier breast compared to the denser tissue.

Figure 1

Why Your Breast Density Matters (Dense Breasts Canada, 2022)



Note. Image A displays the results of a mammogram performed on relatively fatty breast composition. The mass is in image A is easily identifiable. However, in images C and D, the tumor is more difficult to ascertain due to the denser breast composition.

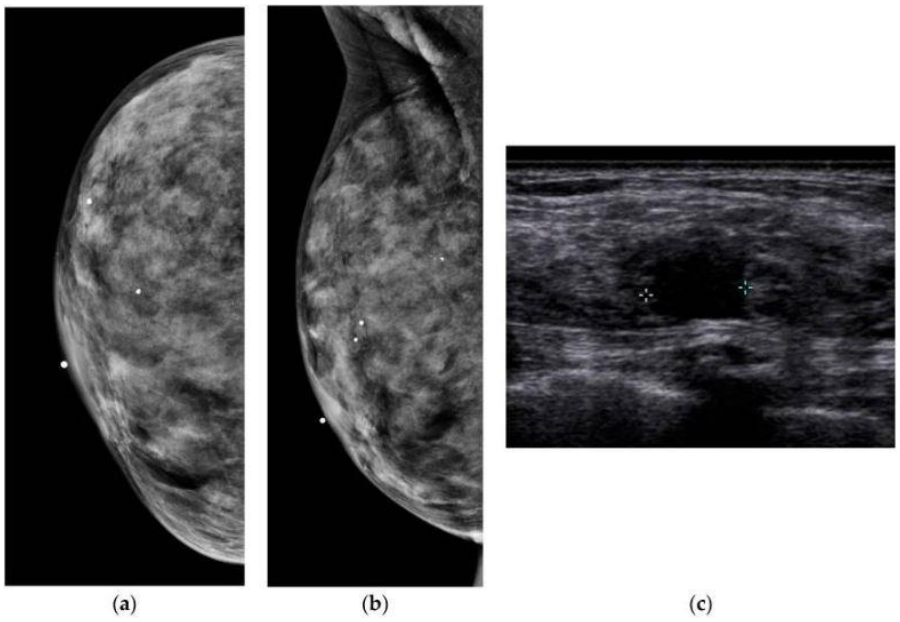
Since roughly half of the population is split between a more dense breast composition vs. a more fatty breast composition (American College of Radiology, 2017), alternative and/or supplemental screening modalities are required.

One common supplemental technique to the mammogram is the ultrasound. A study conducted by Chae et al. (2013) revealed that supplementing ultrasound to a mammogram screening, specifically in women with dense breasts, increased the cancer detection yield by a rate of 2.391 per 1000 women. Additional studies by Kaplan (2001) and Thigpen et al. (2018) also point toward the importance

ultrasound plays in the detection of breast cancer. This is partially due to the fact that cancerous masses have a hypoechoic characteristic (i.e., do not reflect sound waves), whereas the surrounding dense tissue is echogenic (i.e., does reflect sound waves). Figure 2 below contrasts how the occult mass is hidden from the mammogram scans in images (a) and (b), whereas it's readily apparent via the ultrasound scan in image (c).

Figure 2

Mammography vs. Ultrasound (Thigpen et al., 2018)



Note. Mammogram scans in images (a) and (b) hide the existing mass in a dense tissue sample. An ultrasound scan in image (c) easily reveals the mass (flanked by calipers).

Since ultrasound technology has been proven as an effective modality in the early detection of breast cancer, particularly for women with dense breast tissue, the project herein focuses on this common modality. In a joint partnership between Mayo Clinic Health System and the University of Wisconsin – La Crosse, this initiative aims to develop a state-of-the-art computer-assisted detection (CAD) system to aid in the interpretation of breast ultrasound (BUS) images.

Terminology & the Interpretation of Breast Ultrasound Imagery

As the use of medical imaging, particularly mammography, expanded In the 1980s, clinical interpretations of patient images often lacked uniformity, and communication between care providers was inconsistent (Burnside et al., 2009). In the late 1980s, the American College of Radiology (ACR), along with individuals from various professional medical organizations, gathered with the objective of correcting these inconsistencies and address the standardization issues. The result of their effort was the development of the Breast Imaging Reporting and Data System (BI-RADS) atlas. According to the ACR (n.d.), "the BI-RADS atlas provides standardized breast imaging terminology, report organization, assessment structure and a classification system for mammography, ultrasound, and MRI of the breast." The BI-RADS atlas gives radiologists a consistent and clear way to communicate results and helps provide specific management recommendations to patients and other medical providers. See Table 1 for the BI-RADS categories, assessments, likelihood of malignancy, and recommended management.

Table 1

BI-RADS Assessment Categories and Management Recommendations (ACR, n.d.)

Category	Assessment	Likelihood of Malignancy	Recommended Management
0	Incomplete	N/A	Recall for additional imaging
1	Negative	Essentially 0%	Routine screening
2	Benign	Essentially 0%	Routine screening
3	Probably Benign	> 0% to ≤ 2%	Short-interval follow-up and continued surveillance
4 4a 4b 4c	Suspicious for Malignancy Low Suspicion Moderate Suspicion High Suspicion	> 2% to ≤ 95% > 2% to ≤ 10% >10% to ≤ 50% >50% to ≤ 95%	Tissue diagnosis
5	Highly Suggestive of Malignancy	≥95%	Tissue diagnosis
6	Known Malignancy	100%	Surgical excision when clinically appropriate

Note. The BI-RADS category assignment results in the largest inter-observer variability among all features of the breast imaging report. The most difficult to ascertain is observations in categories 3 & 4a.

In addition to the categories listed in Table 1, the BI-RADS atlas also provides terminology to describe various clinical characteristics of any mass uncovered during the imaging procedure. Table 2 lists these additional characteristics for the ultrasound modality, as well as the various class values they can assume. Since the BI-RADS category along with these additional features play such an important role in the eventual outcome and recommended management for a patient, they will be important data points to consider as the project team approaches the development of the CAD system.

Table 2

BI-RADS Descriptive Mass Characteristics for Breast Ultrasound Images

Characteristics	Classes
Shape	Oval, Round, Irregular
Orientation	Parallel, Not Parallel
Margin	Circumscribed, Not Circumscribed, Indistinct, Angular, Microlobulated, Spiculated, None
Echo Pattern	Anechoic, Hypoechoic, Isoechoic, Hyperechoic, Complex Cystic and Solid, Heterogeneous, None
Posterior Features	No Posterior Features, Enhancement, Shadowing, Combined Pattern

Statement of Problem

The field of medical imaging has come a long way since its inception. Not only has it benefited from standardization in terminology and reporting as was discussed in the previous section, but the imaging technology has also greatly improved over the last few decades. New features in ultrasound technology like volumetric ultrasound, elastography, and automated breast ultrasound (ABUS) are paving the way to continuously improve clinical outcomes (Peconic Bay Medical Center, n.d.).

However, despite the improvements in the field, interpretations of breast ultrasound images still rely on the judgment and experience of a well-trained radiologist to spot, assess, and define mass characteristics (Jakubowski et al., 2012). The problem then becomes a classic case of human subjectivity, as characteristics assigned can vary from patient to patient, and radiologist to radiologist. An internal

study performed at Mayo Clinic found that the percentage of positive breast ultrasound biopsies varied in range from 31% to 51%. This means many patients undergo unnecessary medical procedures and experience a non-uniform standard of care, which ultimately leads to increases in medical costs. Of particular importance in this study, is the difficulty in discerning the boundary between the BI-RADS category 3 and 4a, which equates to either a management plan of continued surveillance or a tissue biopsy.

Purpose of the Study

Over the last two decades, the field of artificial intelligence (AI) has experienced tremendous growth. Advancements in computing power alongside the big data revolution have paved the way for groundbreaking research that has propelled the field forward at an extraordinary rate (Fumo, 2018). Therefore, with an emphasis on model explainability, the purpose of this study is to assess the use of artificial intelligence to aid radiologists with the interpretation of lesions discovered in breast ultrasound images during routine clinical screenings.

A sub-field of AI called deep learning, specifically the convolutional neural network (CNN) architecture, is particularly well-suited for working with imagery. Research into the application of deep learning for breast cancer screening has shown great progress in recent years. A study conducted by Shen et al. (2021) found that the application of a CNN to breast ultrasound images decreased false positive rates by 37.3% and reduced requested biopsies by 27.9%, all while maintaining the same level of sensitivity. Another study by Zhang et al. (2016) found that their model could achieve an accuracy of 93.4% when predicting mass pathology, with a sensitivity of 88.6% and a specificity of 97.1%. With so many studies producing results that meet or exceed human-level performance, these tools must segue from research in the lab to application in the clinic.

Significance of the Study

As deep learning began to make its way into many of today's technologies, one of its more troubling attributes has gained wider attention, model explainability. Most state-of-the-art deep learning architectures contain many complex hidden layers, and the number of parameters can range in the millions or even billions; see GTP-3 (Romero, 2021). While all these parameters and complexities in architecture allow deep learning models to generalize well toward their task, it leaves their inner workings somewhat of a mystery. This is acceptable in some settings, however, transparency and explainability are key requirements in the medical imaging domain (Gulum et al., 2021; Singh et al., 2021).

This project emphasizes model explainability for BUS interpretation by including the key clinical characteristics as described in the terminology section of this chapter. By including these characteristics as model outputs, radiologists will have a richer understanding of the CAD's pathological prediction. Furthermore, the study aims to produce a continuous value for the likelihood of malignancy such that radiologists will have a deeper understanding of the BI-RADS category output by reporting precisely where on the scale the CAD suggests. Additional explainability techniques will also be explored, such as the saliency maps to visualize regions of interest in the BUS imagery as described by Shen et al. (2021).

Project Outline

To achieve the goals and objectives of this project, stakeholders from the Mayo Clinic and the University of Wisconsin – La Crosse have agreed upon three distinct phases. The three phases are as follows:

1. Begin the initial model-building process by conducting research into various architectures, methodologies, and techniques that have been empirically shown to produce high levels of performance for BUS interpretation. Items include CNNs, transfer learning, image segmentation,

multi-task learning, image augmentation, and hyperparameter tuning. Candidate models will be interpreted and presented to stakeholders, and only models that meet or exceed the target metrics will move into the next project phase.

2. Improve candidate model[s] and make them scalable so that they can be trained on very large datasets. Additionally, allow model[s] to ingest additional patient metadata, such as age, pre-test probabilities of breast cancer, etc., to enhance performance and model robustness.
3. Build a practical application that can be accessed from ultrasound units and radiologists' clinical workstations to produce automated BUS assessments in real or near real-time.

Limitations

Due to the large scope of designing and implementing a functional CAD system, the research team has been assigned responsibilities for different aspects of the project. The research within this project is limited to activities listed in phases I and II of the project outline section, particularly on deep learning. Therefore, the following chapters in this document will focus on the process of model building, training, evaluation, and scalability.

Chapter 2: Literature Review

Introduction

As was discussed in the previous chapter, the ultrasound modality plays an important role in breast cancer diagnosis since it has various advantages when compared to mammography. For example, it is particularly useful when the patient's breast tissue is radiographically dense (Boyd, 2007; Thigpen et al., 2018). Additionally, ultrasound equipment is relatively cost-effective when compared to mammographic equipment and is more widely available (Berg, 2015). Finally, ultrasounds are favored in scenarios where a patient is deemed at high risk of radiation exposure such as in instances of pregnancy (National Cancer Institute, n.d.). However, despite these advantages, radiologists' interpretations of BUS

images vary widely, and, therefore, have been shown to have a higher level of, or can increase, false-positive biopsy rates (Abdullah et al., 2009; Lazarus et al., 2006).

In an attempt to reduce these levels of intra and inter-observer variability, researchers have proposed various machine learning CAD systems in the past (Chen & Hsiao, 2008; Ding et al., 2012). The ability to automate the interpretation of BUS images provides a promising opportunity to reduce medical costs, avoid unnecessary procedures, and above all else improve patient outcomes. While these past proposals were shown to perform well under the parameters of their study, they relied heavily on hand-crafted features which have proven difficult to generalize across studies with different parameters and datasets (Shen et al., 2021).

Recent advancements in deep learning have shown state-of-the-art performance across many different domains (Pham et al., 2021; Szegedy et al., 2015), and are suited particularly well for computer vision tasks. The literature review that follows focuses on the use of deep learning, specifically convolutional neural networks (CNNs), to aid in the interpretation of breast ultrasound images. It also explores various techniques for working with small datasets, model explainability, as well as different regularization methods.

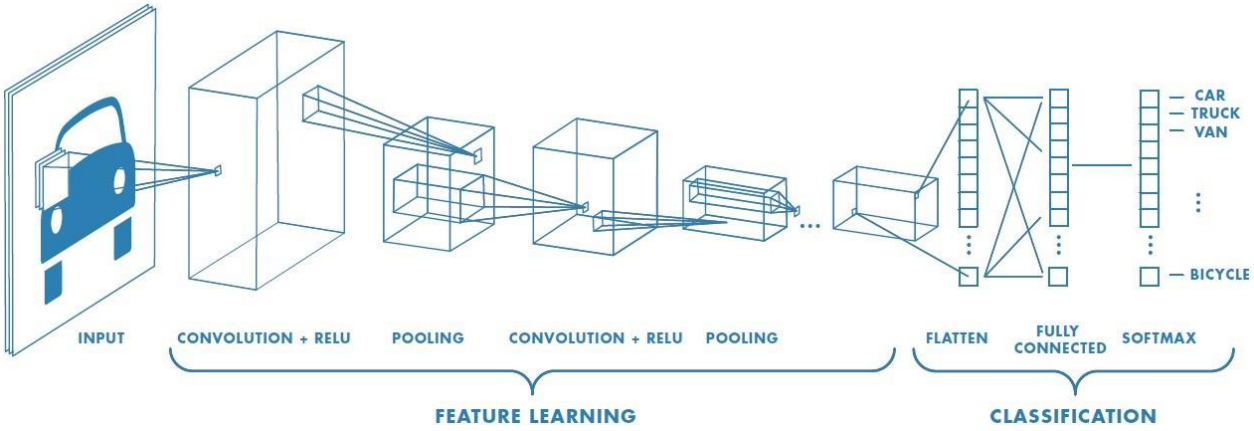
Convolutional Neural Networks

A convolutional neural network is a specific type of neural network that uses a mathematical operation called a convolution within its hidden layers (Goodfellow et al., 2016). In its simplest form, the convolution operation applies what is known as a “filter” to input data to not only capture the most relevant features at that layer but at the same time reduce the number of outputs that move into the next layer of the network. Typically, this is then followed by an activation and pooling operation, which is yet another application of filtering for feature selection and dimensionality reduction. Figure 3

displays the general structure of the CNN architecture. Notice how the inputs are reduced from layer to layer so that only the most relevant features are emphasized as the data moves through the network.

Figure 3

The Basic Building Block of the Convolutional Neural Network (Saha, 2018)



While the idea of the convolution operation in mathematics is nothing new, its application in the field of machine learning began roughly 30 years ago. In their seminal work, LeCun et al. (1989) proposed the idea of the convolutional neural network as a way to work with highly dimensional data, such as imagery, with minimal pre-processing effort. The results of their research yielded one of the very first CNN architectures called LeNet. Their paper showed that by allowing the network to self-learn the most appropriate features, the CNN could outperform models that were based on hand-crafted features by domain experts. Almost 10 years later, LeCun et al. (1998) released LeNet5, which powered some of the first optical character recognition (OCR) systems that automatically read and processed millions of paper checks in the banking industry. Although modern-day CNN architectures such as ResNet, VGG, or DenseNet might look different than their ancestor, they all rely on this common convolution operation to automatically discover the most appropriate domain-specific features and reduce dimensionality.

Since CNNs handle high dimensionality so well, they have held a dominant position in the domain of computer vision where features, i.e., individual pixels, can number in the thousands or even

millions. A large amount of literature exists on the use of CNNs in the field of medical imaging (Sarvamangala & Kulkarni, 2021; Shen et al., 2017), and many of these studies have produced results that either meet or exceed expert-level performance within BUS interpretation (Shen et al., 2021; Zhang et al., 2016). Given the current success of CNNs in classifying medical imagery, it is the primary architecture of choice for this project.

Transfer Learning

Deep learning models are notorious for requiring a lot of data to train on to generalize well in real-world scenarios (Sun et al., 2017). State-of-the-art models such as ResNet, VGG, or DenseNet are trained using a popular open-sourced image database called ImageNet, which contains more than 14 million images across 20,000 categories. Unfortunately, many task-specific datasets are multiple orders of magnitude smaller than ImageNet, and the ability to collect more data can be very difficult and costly. Without enough training samples, these models are unable to learn the most appropriate features for their given task and therefore do not perform well in real-world scenarios. Nowhere is this paradigm more true than in the healthcare industry. With regulations such as the Health Insurance Portability and Accountability Act (HIPAA), procuring a medical imaging dataset at scale is a difficult endeavor. Kohli et al. (2017) cite “data starvation” as the number one challenge for the application of deep learning in the medical imaging field.

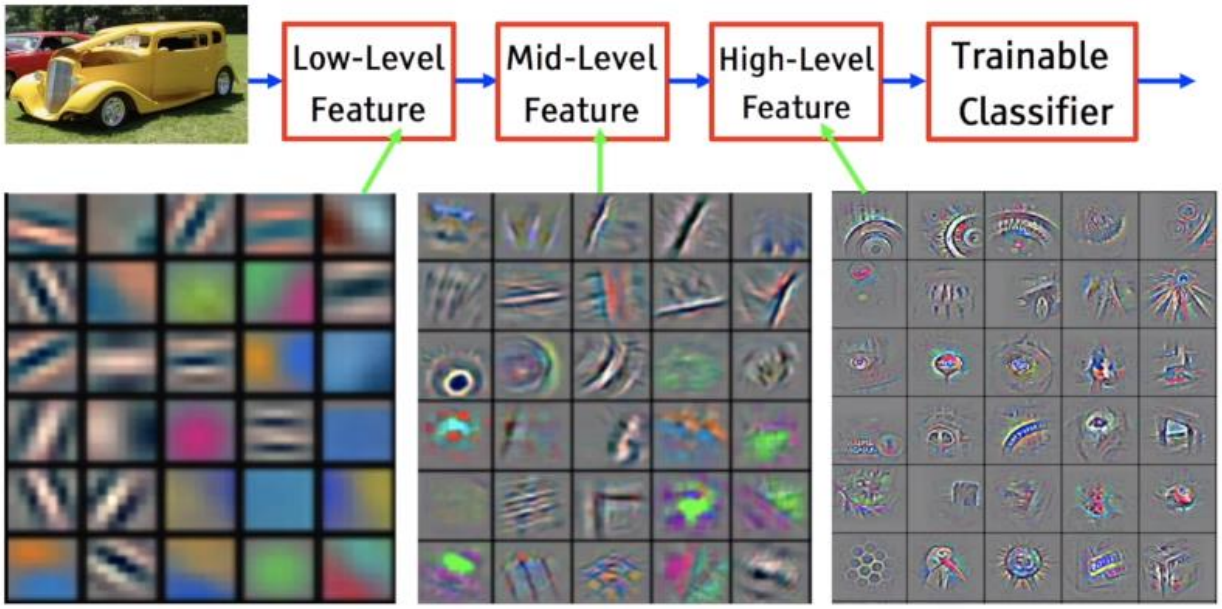
Various strategies have been devised to work around the issue of small datasets. While there is no one-size-fits-all approach, one such technique that has been widely adopted in the field is known as transfer learning. In summary, transfer learning can be defined as taking the parameters (weights) and structure of an existing pre-trained model to use as a starting point for the task-specific problem at hand. These pre-trained models are then adapted for a new target task through a process of fine-tuning (Huh, 2016). Weiss et al. (2016) use the analogy of a scenario where two individuals are learning to play

the piano. Since one individual has experience playing another instrument, they will learn the piano faster than the other individual who possesses no musical experience whatsoever.

One reason transfer learning works well in practice is due to what the CNN learns when training on a dataset. As images move through the network, each layer of the model learns a particular set of features, from low-level items such as lines and edges to mid-level items such as partial shapes. Finally, later layers in the network build on these learned features from the previous levels and learn to detect complex shapes such as a wheel or the frame of a car. Figure 4 provides a visual representation of this learning process, from low to high-level features.

Figure 4

A Visual Representation of the CNN Learning Process (Vignesh, 2020)



Transfer learning capitalizes on the idea that the way CNNs learn basic lines, edges, or fundamental shapes does not change across domains. A line is a line, a circle is a circle: These are fundamental truths of the physical world we live in. Therefore, models that are trained using very large datasets, such as ImageNet, gain a very robust understanding of how these low to mid-level features are represented in our physical world and perform well when transferred across domains (Donahue et al., 2013).

The use of transfer learning is well-studied in the field of medical imaging. A review of the literature by Morid et al. (2021) surfaced over 102 studies on transfer learning. Specifically, Byra (2021) was able to achieve a BUS lesion classification accuracy of 91.5% using transfer learning. With such a rich body of evidence supporting transfer learning, it becomes a key factor in the success of this project due to the small dataset currently available from the Mayo Clinic.

Data Augmentation

Data augmentation is another important technique used to combat the problem of small datasets (Shorten & Khoshgoftaar, 2019). One type of augmentation strategy is to stochastically apply geometric transformations to images during the model's training phase. These include methods such as flipping the image along an axis, rotating it by any number of degrees in either direction, cropping, padding, squishing, applying noise reduction filters, zooming in or out, etc. Any or all methods can be applied to a single image at training time, and are typically done so at random. Figure 5 shows common a set of geometric augmentations applied to an image of a dog.

Figure 5

Commonly Used Image Data Augmentation Techniques (Alto, 2020)

The theory behind why geometric augmentation works so well is that it provides the model with additional examples of what it could likely expect to encounter in the real world: subjects from different viewpoints, scales, occlusions, levels of noisiness, background variations, etc. (Shorten & Khoshgoftaar, 2019). These geometric methods have been shown to improve the performance of deep learning models within the field of medical imaging. A study by Hussain et al. (2017) showed that methods such as the flip, and the application of Gaussian filters, to a sample of randomly selected medical images from the Digital Database for Screening Mammography (DDSM) lead to validation accuracies of 84% and 88% respectively. Considering the overwhelming evidence in support of geometric data augmentation methods within the field of computer vision, these will become important strategies to explore throughout this project.

Regularization Techniques

A problem that every deep learning practitioner encounters at some point along the way is the issue of overfitting. A model is considered to be overfitted when its learned parameters align too close to the sample data that was used during training (IBM, n.d.). Overfitted models cannot generalize well against new unseen data, and, therefore, are not useful at inference time. Researchers have proposed many different techniques, collectively called regularization, to combat the problem of overfitting. Techniques such as batch normalization (Ioffe & Szegedy, 2015) and dropout (Srivastava et al., 2014) have become best practices in the field of deep learning. Both batch normalization (Ezzat et al., 2021), and dropout (Roth et al., 2015) have been utilized with success in the field of medical imaging, and therefore will be investigated for this project.

Another regularization technique that has gained recent attention is called mixup (Zhang et al., 2018). In summary, mixup is a data augmentation technique with a regularization effect that linearly

synthesizes two observations at random from the training data. The resulting convex combination is then used in the training data set. Mathematically, Zhang et al. (2018) define mixup as:

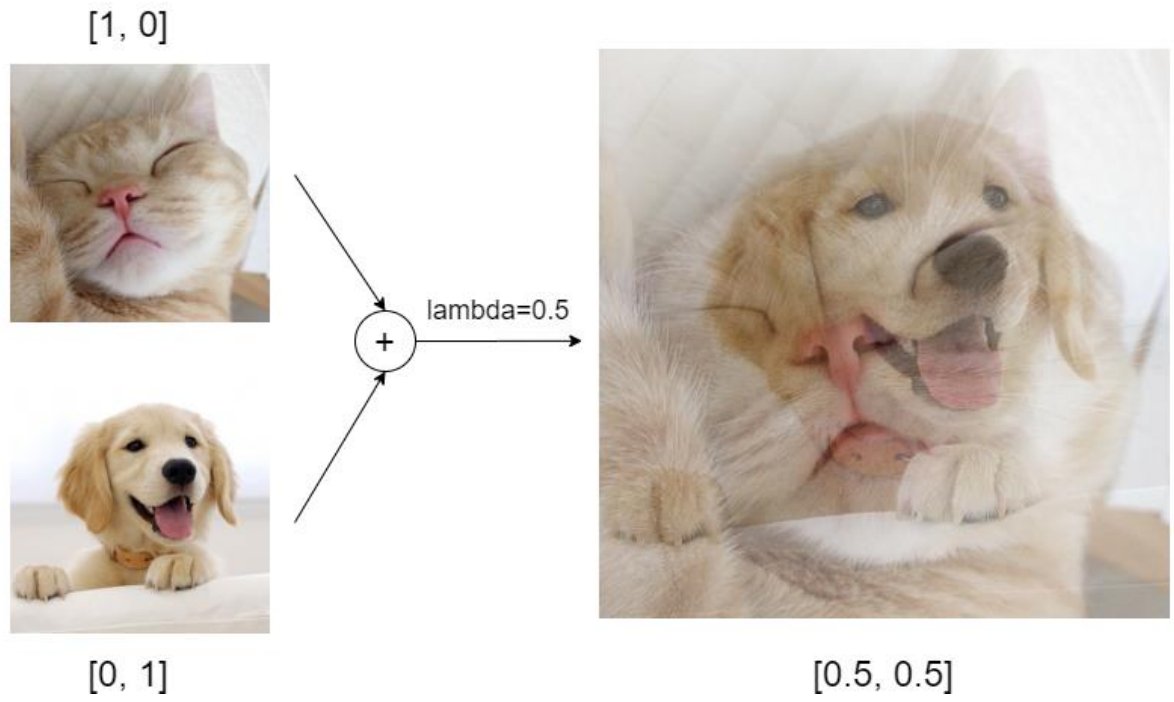
$$x' = \lambda x_i + (1-\lambda)x_j \quad \text{where } x_i \text{ and } x_j \text{ are raw input vectors}$$

$$y' = \lambda y_i + (1-\lambda)y_j \quad \text{where } y_i \text{ and } y_j \text{ are one-hot encoded label vectors}$$

(x_i, y_i) and (x_j, y_j) are the two randomly sampled observations from the training data, and $\lambda \in [0,1]$ where $\lambda \sim \beta(\alpha, \alpha)$. Figure 6 shows an example of mixup being applied to a computer vision problem.

Figure 6

Synthetic Sample Created Using the Mixup Technique (Khatua, 2020)



Note. The images are linearly combined at a factor of $\lambda = 0.5$ *resulting in the convex combination of* x'*. Subsequently, the resulting one-hot encoded label vector* y' *is set to [0.5, 0.5] cat and dog respectively.*

Not only does mixup provide more unique training data through its random synthesis of x', but it also makes it harder for the model to memorize particular features of the target class[es] and succumb to the issue of overfitting. Mixup has been shown to be a useful technique for training deep learning

models in the field of medical imaging (Galdran et al., 2021). A study by Wang et al. (2020) showed an improvement of 17.3% over state-of-the-art results using a variant of mixup they called FocalMix. Since this technique has demonstrated powerful results and is particularly useful with smaller datasets, it will be explored during this project.

Multi-task learning

Multi-task learning (MTL) is an approach in deep learning that simultaneously trains a model on a set of related tasks. MTL can be accomplished by using one of two different strategies. Hard parameter sharing allows the model to share a set of common parameters that are jointly optimized against the loss function. Conversely, soft parameter sharing allows each task its own set of parameters, but a set of regularizations is applied to encourage these parameters to become more similar to each other. The motivation for MTL is that in theory, the features learned for each related task can help the model better learn the other tasks through what is known as inductive transfer (Caruana, 1997). Ruder (2017) provides an analogy stemming from biology, where a newborn first learns to recognize faces and then applies this knowledge to recognize other objects such as a bottle or a toy.

The benefit of an MTL approach is two-fold. First, MTL helps with the prevention of model overfitting. It increases generalizability since the model must optimize in a way to best represent all related tasks simultaneously. Secondly, MTL provides a mechanism to embed the notion of explainability directly into a model's parameters as it trains, whereas other approaches to explainability are post-hoc in nature and may not reflect the model's actual reasoning (Chen et al., 2019).

MTL models are often set up in a task-hierarchical fashion, where a series of task outputs can be used as input for the next, sometimes more complex, task. Nowhere is this task-hierarchical approach more beneficial than in the field of healthcare, where a medical professional takes multiple points of empirical evidence to make a diagnosis. As was discussed in Chapter 1, the BI-RADS lexicon provides

radiologists with a common language to describe and interpret BUS lesions. The radiologist first defines the various clinical characteristics as noted in Table 2. Next, the resulting evidence is used to make predictions on the BI-RADS category and the recommended management plan as defined in Table 1. MTL provides a technique that allows a deep learning model to mimic the process and logic employed by radiologists to interpret BUS images. Finally, MTL is well supported in the medical imaging literature (Marques et al., 2021; Wang et al., 2022), and therefore it will be explored as a strategy to not only improve model performance but also as a way to provide explainability to our stakeholders.

Summary

This chapter provided an opportunity to explore the literature in regard to deep learning within the field of medical imaging. Various approaches to overcome the challenges of a small dataset were a key focus of this literature review, as the current dataset from Mayo Clinic is roughly 234 images in total. While many of these techniques can be applied generically across domains, evidence of their impact on the medical imaging field was discovered in the literature.

Another focus of this chapter was on model explainability. Through the use of the multitask learning (MTL) strategy, explainability is embedded directly into the model's parameters during training since parameter sharing forces a model to simultaneously optimize performance for various relevant tasks. MTL also supports a task-hierarchical learning structure which allows the outputs of specific tasks to act as inputs into another task. Task-hierarchical learning can be utilized as a way to mimic the real-world process of a radiologist interpreting BUS images, which further builds understanding and trust within the CAD system. The next chapter of this paper focuses on the model building process, and will apply knowledge acquired through the review of the literature.

Chapter 3: Methodologies

Introduction

Since the purpose of this study is to assess the use of deep learning to aid radiologists in the interpretation of BUS images, this chapter is dedicated to the overall methodologies that were used to achieve this goal. It contains information about the dataset used to train the deep learning models, the multitask approach to model architecture, and the training process with the various techniques used to improve performance as explored in the previous chapter.

Dataset & Preparation

Medical imaging data is almost universally stored in a file format known as Digital Imaging and Communications in Medicine (DICOM). Although DICOM is a protocol that addresses much more than data storage, DICOM files typically contain both images and other metadata about the patient and study. Examples of patient metadata include items such as patient name, patient identification number, date of birth, etc., whereas study metadata are items such as study date, imaging modality, image resolutions, machine model, etc.

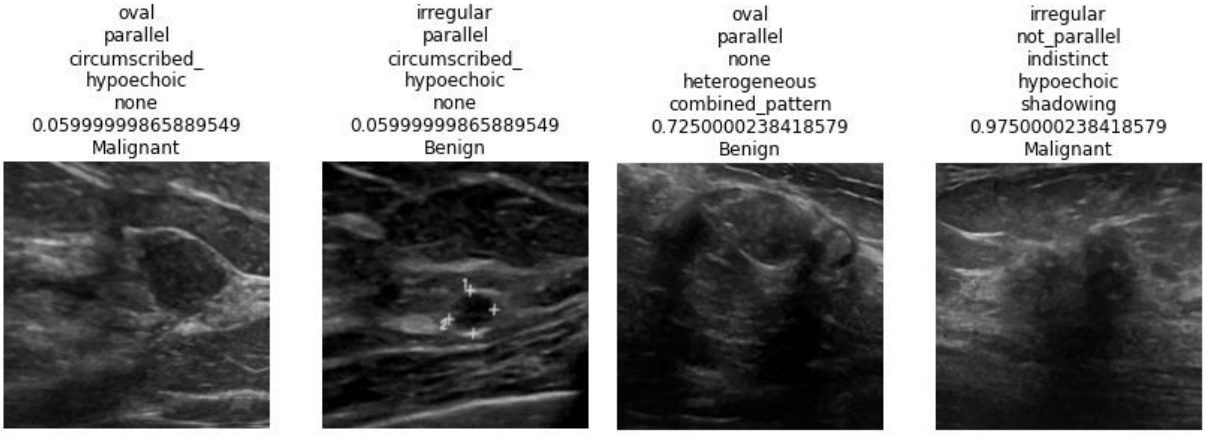
For this project, stakeholders at the Mayo Clinic and the University of Wisconsin – La Crosse identified and secured a small sample of patient studies in the DICOM format. Mayo Clinic's data privacy team applied a robust anonymization process to the DICOM files to comply with HIPAA and other federal privacy regulations. One part of the de-identification process was to replace the patient's identification number with a randomly generated value. Additionally, annotation markings containing personally identifiable information (PII) around the borders of the images were removed. Further steps were taken to clean the data once it was released, such as applying another round of cropping to the images to extract the region of interest using planar coordinates supplied within the DICOM files. It should be noted that some images still retained additional noise artifacts such as caliber markings that

flank the lesion. At the time of this writing, efforts were underway by other team members to explore additional image cleaning techniques, such as inpainting (Xie et al., 2012), to fully remove the remaining image artifacts.

Images were saved in the .PNG format with a file name comprised of the de-identified patient ID number, pathological result (benign or malignant), and the angle of the ultrasound transducer (longitudinal or transversal). The final images were 3-channel RGB, with an average resolution of 565×565 pixels. In total, 234 images across 126 unique patient IDs were included in the dataset. Figure 7 displays a sample of the images and their class labels.

Figure 7

Sample of Training Images from the Mayo Clinic Dataset



Note. BI-RADS labels were converted to a continuous value that represents the median of the likelihood of malignancy ranges per the ACR's BI-RADS atlas. This approach provides radiologists with a more precise understanding of model inference.

Stakeholders from the Mayo Clinic provided annotations for each image in .CSV format. The images were matched to their respective labels via the de-identified patient ID number, which was both available in the .CSV file and image file name. Minor data clean-up steps were required once the annotations were matched to their images. These included the following steps:

- Setting “Elevated Risk” to “Benign” in the pathology variable and removing any rows with an “Unknown” value
- Removing any rows with a missing value for the BI-RADS score
- Removing any rows missing a BUS image

Additionally, BI-RADS scores were converted to a continuous value representing the median of the likelihood of malignancy ranges as defined in the ACR’s BI-RADS atlas; see Table 1. This approach provides end-user radiologists with an additional layer of explainability as they understand precisely where on the likelihood of malignancy scale the model made its prediction.

As a final step in the data pre-processing phase, the fully labeled images were split into two subsets, one for model training, and the other for model validation. At the time of this writing, 108 of the 126 patients within the dataset had two images representing both the longitudinal and transversal views respectively. Therefore, the dataset was stratified between those patients with multiple images vs. those with a single image. This strategy prevented a single patient from being included in both the training and validation datasets simultaneously and avoids the potential for data leakage during training. Finally, a randomized split using the proposed stratification approach was employed to sample roughly 80% of the images for the training set, and the remaining 20% were placed into the validation set. Five-fold cross-validation using the aforementioned method was applied during the model training process. Five-fold cross-validation makes use of the limited data available in the Mayo Clinic dataset, and provides a more accurate assessment of model performance.

Class Imbalances

A common problem with real-world datasets is that their observations are generally skewed toward one particular class or set of classes. A model may show great results on a metric like accuracy, however, a closer inspection of the dataset typically reveals a high level of imbalance between classes.

In practice, the model maximizes its performance by predicting the majority class for every observation and the minority classes are ignored and often go unlearned. There are many techniques to combat the effects of an imbalanced dataset, such as combining minority classes or the application of class weights into the loss function, to oversampling and more elaborate techniques like SMOTE (Chawla et al., 2002).

For this project, exploratory data analysis uncovered that 3 of the 6 classification features were highly imbalanced in the prepared dataset provided by Mayo Clinic. Table 3 displays the largest and smallest classes for each feature, along with a calculated imbalance ratio. Shape, Margin, and Echo Pattern were the most imbalanced.

Table 3

Feature class frequencies and imbalance ratios of the Mayo Clinic dataset

Feature	Largest	Smallest	Ratio
Shape	Irregular 114	Round 8	14.25
Orientation	Parallel 145	Not Parallel 90	1.61
Margin	Circumscribed 97	Not Circumscribed 4	24.25
Echo Pattern	Hypoechoic 175	Hyperechoic 2	87.5
Posterior Features	Shadowing 91	Combined Pattern 20	4.55
Pathology	Malignant 123	Benign 111	1.10

Note. There is no official definition of when a feature is "too" imbalanced, however, balancing techniques are an effective strategy when the imbalance is extreme such as in the case of the Mayo Clinic dataset.

The pros and cons of the three balancing strategies mentioned above were considered when addressing the imbalance issue in the Mayo dataset. It was determined that combining extreme minority cases into an "other" category would not be in alignment with the explainability goals of this project. Limiting the model's ability to predict discrete minority classes prevents radiologists from

making a proper assessment of model outputs, and erodes trust in its performance. Oversampling was also considered for this project; however, this technique can lead to model overfitting due in part to the limited data available (Sowjanya & Mrudula, 2022). Therefore, the application of class weights was selected as the strategy to manage the imbalanced dataset. For each class within each categorical variable, a weight was calculated and applied to the loss function during the training process. With class weights in effect, misclassifications in majority classes are penalized less, whereas errors in minority classes are penalized more heavily. The model then tends to "favor" correctly classifying minority classes, and thus improves overall generalization.

A final technique to correct for the imbalanced dataset was explored during the data augmentation phase of the training process. Mixup (Zhang et al., 2018), as described in the previous chapter, linearly combines two randomly sampled images and their respective class labels using a mixing factor λ, where $\lambda \in [0,1]$, and $\lambda \sim \beta(\alpha, \alpha)$. Since class labels of these new synthetically generated images are a linear combination of their original values, class weighting strategies are not as effective when utilizing the mixup technique. To account for class imbalances under the mixup scenario, a technique called remix (Chou et al., 2020) was utilized. Remix is applied in addition to the original mixup technique, and aims to favor minority classes via a relaxation of the original mixing factor λ. Chou et al. (2020) mathematically define remix as the following:

$$\lambda_y = \begin{cases} 0, & \frac{n_i}{n_j} \geq \kappa \; and \; \lambda < \tau; \\ 1, & \frac{n_i}{n_j} \leq \frac{1}{\kappa} \; and \; 1 - \lambda < \tau; \\ \lambda, & otherwise \end{cases}$$

Here the $\lambda_y$denotes that only the class label mixing factor is adjusted under the remix regiment. n_i is defined as the number of samples of the *original* input image's class label from the entire original dataset. n_j is the number of samples of the *randomly* selected image's class label from the entire

original dataset. κ and τ are hyperparameters that allow the user to define the sensitivity of remix. In essence, remix will set the synthetically mixed sample's class label, y', entirely to the minority class if any of the criteria above are met, otherwise it will use the original mixing factor λ when interpolating the labels. Remix has a similar effect to the class weighting scenario as discussed above. The default values suggested by Chou et al. (2020) of $\kappa = 3$ and $\tau = 0.5$ were used for the Mayo Clinic dataset.

Modeling

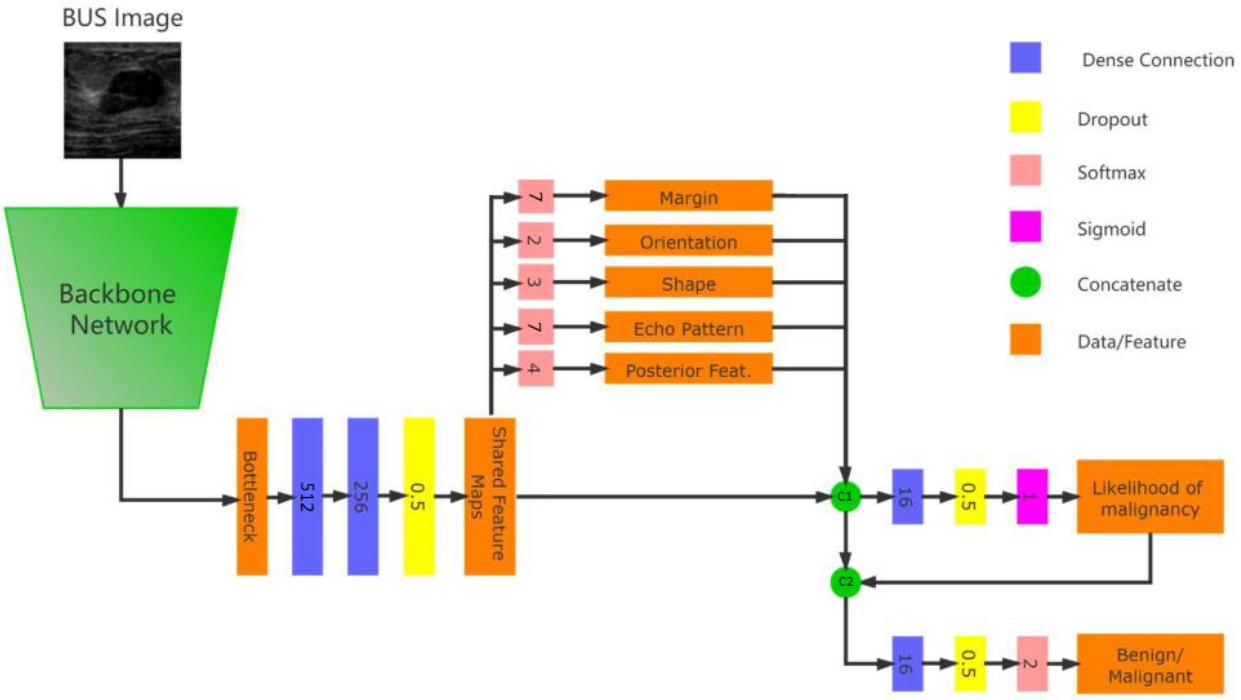
This section is dedicated to the modeling and research explored during the scope of this project. It should be noted that the intent of this project was not to create a novel deep learning architecture for BUS interpretation. Instead, the scope of this project was focused on the application of state-of-the-art models, and exploring the latest techniques proposed in the literature on CAD systems for BUS interpretation. Therefore, a great deal of inspiration was derived from a paper titled *BI-RADS-Net: An explainable multitask learning approach for cancer diagnosis in breast ultrasound images* authored by Zhang et al. (2021). Although the model described in this section closely follows the proposed BI-RADS-Net architecture by Zhang et al., modifications were made to align with the Mayo Clinic dataset. Additional differences occurred during the data pre-processing and training phases, which are noted throughout this chapter.

As was discussed in chapter 2, multi-task learning (MTL) is an approach to model architecture that simultaneously trains a model on a set of related tasks using either a hard or soft parameter sharing technique. A benefit of the MTL approach was that it helps prevent model overfitting since the model optimizes for all features simultaneously. Secondly, the approach provides a mechanism to embed explainability directly into a model's parameters as each learned task can serve as input towards a subsequent task. This hierarchical learning can be thought of as mimicking the approach radiologists follow to make a diagnosis from a BUS image, therefore enhancing the explainability of the system.

Figure 8 displays a visual representation of the modified BI-RADS-Net multitask architecture employed within this project.

Figure 8

Modified BI-RADS-Net architecture adapted for Mayo Clinic dataset, originally proposed by Zhang et al. (2021)



The BI-RADS-Net architecture begins with using a pre-trained state-of-the-art model, known as the backbone network. For this project, ResNet-34, VGG-16, and DenseNet-201 were all tested as backbone networks as they are all popular, well-studied, state-of-the-art models with a strong pedigree. A model head was created containing layers of pooling, batch normalization, dense connections, ReLU activations, and dropout. Output from the backbone network was passed through the model head, resulting in the shared feature map as displayed above in Figure 8. Individual classification tasks for each BI-RADS descriptor were attached to the shared feature map via a simple linear connection. The output of each of these short classification branches was the predicted descriptor class, as determined by the softmax function.

Next, the predicted BI-RADS descriptors were concatenated with the original shared feature map, labeled C1, and were used as input into the regression branch for the likelihood of malignancy. A single dense layer, followed by dropout, was used before obtaining a continuous value prediction via the sigmoid function.

Finally, the model predicts the observation's pathology. It used the concatenated output that fed into the regression branch from the prior step, C1, and joined it with the output from the likelihood of malignancy. This concatenation, C2, was then fed into the pathology branch through a dense layer, dropout, and the final SoftMax layer to receive the predicted pathology.

One main architectural difference exists between the original BI-RADS-Net proposed by Zhang et al. and the modified version described above. In BI-RADS-Net, the margin descriptor contains only two classes, circumscribed and not circumscribed, and is then broken out into four distinct sub-classes for angular, microlobulated, indistinct, and spiculated. These four sub-classes are binary, meaning the observation either has the characteristic or it does not. The Mayo Clinic dataset was not structured in this fashion; therefore, all margin classes fell under the marginal zone descriptor.

Training

The final section of this chapter discusses the approach used to train the modified BI-RADS-Net model. It contains information about the data augmentations performed at runtime, the customized loss function for the multi-task approach, finding and selecting optimal learning rates for the optimizer, and the metrics used to gauge overall model performance.

As was discussed in chapter 2, data augmentation techniques are an essential part of training a practical deep learning model. This is especially true when a dataset contains limited data as is the case for the medical imaging domain. However, geometric transformations must be considered carefully as they could adjust an image in such a way that alters the final classification label, or worse, generate

something that is considered outside the scope of the problem domain. Therefore, careful consideration went into the geometric transformations used to alter the BUS images. For this project, both individual images and image mini-batches had transformations applied at training runtime using the graphics processing unit (GPU). See Table 4 for a complete list of geometric transformations used in this project.

Table 4

A List of geometric transformations applied during the training process

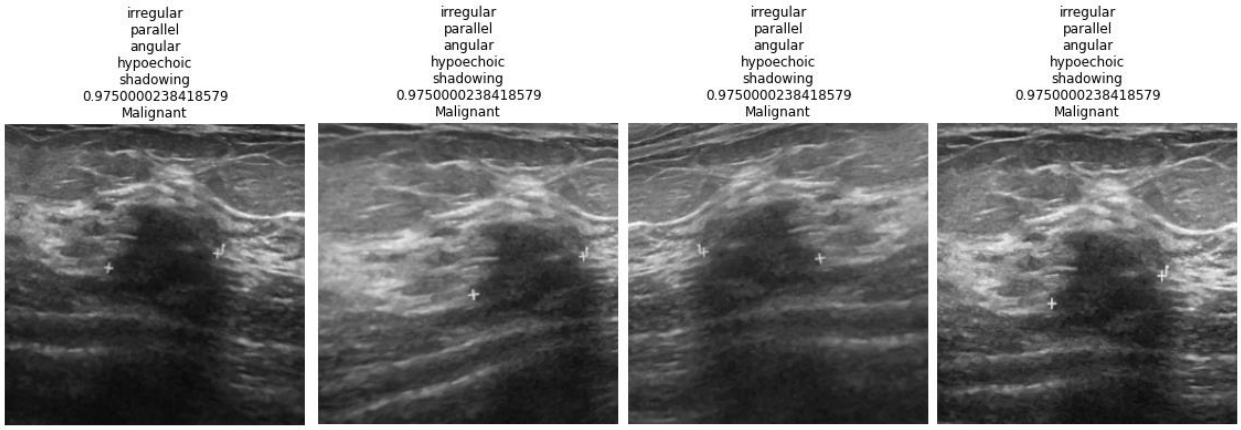
Scope	Method	Value	Effect
Item	FlipItem	0.50	Randomly flip an image along its vertical axis using a probability of 0.50.
	RandomResizeCrop	256	Crop a random section of the image, resize to 256x256 pixels, and scaled appropriately.
Batch	Rotate	10°	Randomly rotates an image in the right or left direction by a maximum of 10°.
	Zoom	1.10	Randomly apply a 1.10x maximum zoom.
	Lighting	0.20 0.75	Randomly apply a change in brightness by 0.20, using a probability of 0.75.
	Warp	0.20	Randomly adjusts the perspective of an image by a magnitude of 0.20.

Note. Some augmentations would have brought images outside the domain of BUS imaging and were therefore not allowed at runtime. Values were also restricted to prevent out-of-domain augmentations.

Images were resized to 256x256 pixels to fit the input requirements of the pre-trained state-of-the-art models discussed in the previous section. Images were allowed to be flipped along the vertical axis but not along the horizontal axis as this would create images outside the domain of BUS since ultrasounds are not performed upside down in a clinical setting. Image rotation was allowed; however, to avoid the similar out-of-domain issue as the horizontal flip, a maximum of 10° was allowed in either direction. Finally, slight warping and lighting effects were also allowed, as perspectives of the transducer can slightly vary (in pixel space), and the contrast of BUS images can differ depending on the model of the ultrasound machine. Figure 9 displays a sample of these geometric augmentations applied to a single BUS image.

Figure 9

Geometric Augmentations Applied to a Single BUS Image from the Mayo Clinic Dataset



A custom loss function was created to train the modified BI-RADS-Net. The multi-task model is a mix of six classification tasks in addition to one regression task. Cross entropy was selected as the loss function for the classification tasks, and mean squared error (MSE) was selected for the regression task's loss function. As described in the class imbalance section of this chapter, both class weights and a relaxation of the mixing factor λ under the remix regiment for mixup (Chou et al., 2020) were applied to the classification tasks. MSE was linearly interpolated directly and does not have class weights since it is a regression task. Finally, a weighting was applied to each feature to denote its importance within the loss function. Below is a mathematical description of the loss function used for the modified BI-RADS-Net.

$$J = \sum_{i=1}^{t} \omega_i \mathcal{L}_i(X_i, Y_i)$$

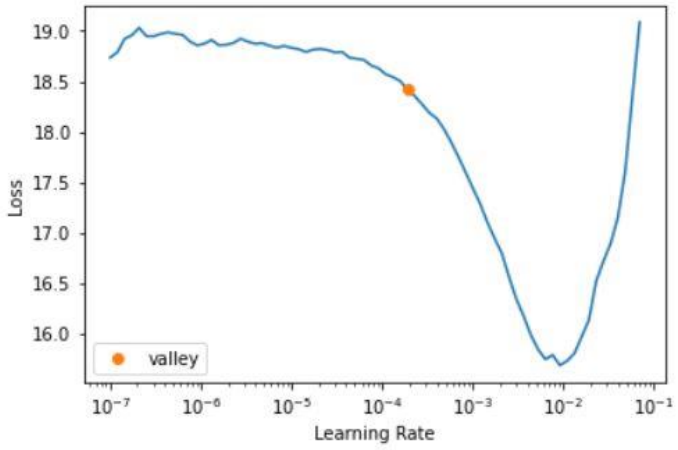
Here ω_i represents the current task's feature importance weight. $\mathcal{L}_i$ is the current task's loss function with the appropriate mixup regiment and class weights (if applicable) as discussed in the previous paragraph. X_i and Y_i are the current task's data. Losses from each of the seven tasks, t, are summed together and the overall cost function J is optimized via gradient descent.

The optimizer known as Adam (Kingma and Ba, 2017) was used to train the modified BI-RADS-Net. Adam is an enhanced version of stochastic gradient descent (SGD), and it combines best properties of both the AdaGrad and RMSProp algorithms making it particularly useful for sparse and noisy problems such as computer vision (Brownlee, 2021). Hyperparameters for Adam were left as default with the exception of the learning rate, which followed a two-step approach.

First, since the modified BI-RADS-Net uses a pre-trained state-of-the-art model as its backbone, only the newly added branches were trained during the initial epochs. This approach specifically targets the parameters of the newly added branches allowing the model to quickly fine-tune their randomized values to the task of BUS imaging. The technique known as the learning rate finder (Smith, 2015) was used to select an optimal learning rate for Adam. The learning rate finder technique works by iteratively training the first mini-batch using a slightly larger learning rate each time to test what occurs to the training loss. The idea is to find the largest learning rate possible before model loss tends to increase exponentially. Figure 10 shows an example of the output of the learning rate finder using the modified BI-RADS-Net. Here it found that a learning rate of slightly larger than 10^{-4} worked best for the initial tuning of the new branches for the first mini-batch.

Figure 10

Learning Rate Finder Applied to the Modified BI-RADS-Net Using a Pre-trained ResNet-34



After the initial fine-tuning occurred on the newly added branches, all model parameters, including the pre-trained ones from the backbone network, are unfrozen and allowed to be trained. For this second round of training, a technique known as discriminative learning rates was applied. Recall from chapter 2 that as a CNN trains, its first few layers learn representations of low-level features such as edges and lines. Knowing that these low-level image features are omnipresent across domains, the parameters from these initial layers of the pre-trained model should require minimal changes from their original values. Discriminative learning allows very small learning rates for initial layers, somewhat larger learning rates for middle layers, and the largest learning rates for the final layers. This technique provides more stability to training entire models under the transfer learning scenario such as the modified BI-RADS-Net used in this project.

Finally, various metrics were used to properly assess the model's performance against the objectives of the project. Accuracy is calculated and monitored for the six classification tasks. Root mean squared error (RMSE) is monitored for the one regression task. To gauge overall performance, a custom metric was created that combines the error rates from each task. Again, RMSE is used for the one regression task, and the error for the six classification tasks was calculated simply as $1 - accuracy$. This combined error was used to select the best model from this study.

Summary

This chapter holistically covered the methodologies used in developing the model for BUS interpretation. First, the data needed to be extracted from the raw DICOM format and cleaned. Next, an exploration of the dataset found it to be imbalanced, therefore various techniques were applied to correct this issue. A review of the literature uncovered a multi-task model architecture, BI-RADS-Net, that was modified for the Mayo Clinic dataset. Finally, a custom loss function, custom metrics, and a variety of performance-enhancing techniques were utilized to train the modified BI-RADS-Net model.

Chapter 4: Results

Introduction

This chapter reviews the results and findings of the models and training procedures discussed in the previous chapter. Models are assessed using all 234 images via a stratified five-fold cross-validation technique. Model architectures are interpreted via a confusion matrix using pathology as the main feature of interest. Additional metrics are calculated and used for the comparison of architectural performance. Finally, the best-performing model architecture is selected and used to draw comparisons against the performance of a trained radiologist.

Before moving on to the model findings section, it should be noted that in the healthcare industry, it is particularly important that malignant cases are identified accurately since the outcome of missing a malignant lesion can have life-impacting consequences. Therefore, one of the main criteria used to gauge model performance in this project is the model's ability to correctly classify a known malignancy (i.e., maximizing true positives), while trying to avoid mistakenly misclassifying a true benign as malignant (i.e., minimizing false positives). These measures are referred to as the model's sensitivity (sometimes called recall) and precision, and they both play an important role in selecting the best model from this study.

Model Findings

To begin the training for each model architecture, the best initial learning rate for each fold was found using the learning rate finder technique (Smith, 2017). Next, the pre-trained parameters are frozen, and the new model branches are trained for 10 initial epochs using the best learning rate found. Now that the new branches are primed for the current task, the pre-trained parameters are unfrozen, and the entire modified BI-RADS-Net is trained for 300 epochs using the discriminative learning rate methodology. Metrics and predictions from each fold of the best model found during the 300-epoch

training are persisted to disk so they can be interpreted later. Finally, model parameters are reset, and the next fold is loaded into the model for training. Once the training process is complete, a confusion matrix is constructed using predictions collected from the best model found during each fold of the five-fold cross-validation process.

The first model trained used the ResNet architecture. ResNet is an important architecture in deep learning since they allow the construction of very deep networks via the use of residual connections (He et al., 2015). Practitioners consider ResNets as one of the standard state-of-the-art models to use in computer vision tasks. Table 5 displays the confusion matrix results using the ResNet-34 architecture as the backbone network for the modified BI-RADS-Net.

Table 5

ResNet-34 Pathology Confusion Matrix

n=234		Predicted Values: Malignant	Predicted Values: Benign
Actual Values	Malignant	97	26
	Benign	41	70

The results in Table 5 show that using the pre-trained ResNet-34 provides a decent level of performance in one of the desired goals; its highest outcome was correctly predicting malignant lesions (true positives), and its lowest score was predicting malignant lesions as benign (false negatives). This combination of outcomes shows an affinity of the ResNet-34 to maximize the model's sensitivity.

Next, a pre-trained VGG-16 model is used in place of the backbone network and trained using the same procedure described above. VGG is considered another standard model architecture used in

computer vision. In contrast to ResNet, VGG uses multiple small convolutional layers in order to extend the depth of the architecture (Simonyan & Zisserman, 2014). Table 6 displays the VGG-16's confusion matrix.

Table 6

VGG-16 Pathology Confusion Matrix

n=234		Predicted Values: Malignant	Predicted Values: Benign
Actual Values	Malignant	103	20
	Benign	56	55

VGG-16 performs slightly better than ResNet-34 with regard to its sensitivity metric. However, the VGG-16 model was roughly split in its ability to predict the benign case, thus leading to a lower precision. A model that performs well with both sensitivity and precision is desired, and a metric that harmonizes the two will be discussed in the next section on model selection.

Finally, a pre-trained DenseNet-201 model is placed into the backbone network of the modified BI-RADS-Net. DenseNets are another important architecture used in practical applications of deep learning. The architecture allows the construction of deep networks by employing all features obtained from all the preceding layers as inputs into the next layers. This approach strengthens feature propagation from layer to layer, encourages feature reuse, and reduces the total number of parameters required to train the model (Huang et al., 2018). Table 7 displays the results of the DenseNet-201's training on the Mayo Clinic dataset.

Table 7

DenseNet-201 Pathology Confusion Matrix

n=234		Predicted Values	
		Malignant	Benign
Actual Values	Malignant	79	44
	Benign	55	56

The DenseNet model did not perform as well as the other two architectures. Both its sensitivity and precision were the lowest of the three architectures explored in this project. These mediocre results could be indicative of limitations within the dataset, which is likely to be the case given the small sample size for the current Mayo Clinic dataset.

This section reviewed the results of three different very popular deep learning architectures that were used as the backbone of the modified BI-RADS-Net. Model predictions were compared to the known pathological outcomes and displayed as a set of confusion matrices. The next section will review the results in greater detail by calculating several performance metrics, and the best model architecture will be selected based on a metric that harmonizes both sensitivity and precision.

Model Selection and Interpretation

As was discussed at the beginning of this chapter, a major priority of any CAD system within the healthcare industry is to ensure that patients with truly malignant lesions are identified since a misdiagnosis can have life-altering impacts. On the other hand, it's also important to avoid misclassifying benign lesions as malignant since this incurs unnecessary surgeries and increases medical costs. F1 score is a commonly used metric to assess the performance and validity of a model in the healthcare setting.

The F1 score works by providing a balance between both sensitivity (i.e., catching truly malignant lesions, while minimizing its errors in this task) and precision, which is the model's ability to predict a benign outcome when the lesion is truly benign while minimizing its errors. F1 score is calculated using the following equation:

$$F1\ Score = \frac{2 * (precision * sensitivity)}{(precision + sensitivity)}$$

Models with high F1 scores can be viewed as providing both a high level of sensitivity and precision and is therefore the metric used to select the best candidate model for application within the CAD system. Table 8 displays the F1 scores and additional metrics for each architecture explored within the scope of this project.

Table 8

Architectural Performance Metrics Using Best Results From Five-Fold Cross-Validation

Pre-trained Architectures	F1 Score	Sensitivity	Precision	AUC	Accuracy
ResNet-34	**0.743295**	0.788617	0.702898	0.709624	0.713675
VGG-16	0.730496	0.837398	0.647798	0.666446	0.675213
DenseNet-201	0.614785	0.642276	0.589552	0.573390	0.576923

The results above show that the ResNet-34 architecture has the highest F1 score and is therefore selected as the best candidate architecture to use in the backbone position of the modified BI-RADS-Net. It should be noted that VGG-16 was roughly one point away from the ResNet-34; therefore, both architectures should be reconsidered as more data is made available to the research team.

Now that the best backbone architecture is selected, its performance under the multitask learning approach can be evaluated. Table 9 shows the ResNet-34's performance for each of the clinical descriptors. Values are reported for each of the five cross-validation folds, and an average of the results

is calculated at the bottom of the table. Accuracy is reported for all descriptors except for the likelihood of malignancy, which uses root mean squared error since it is a regression task.

Table 9

Multitask Learning Results Using the ResNet-34 Architecture

Fold	Shape	Orientation	Marginal Zone	Echo Pattern	Posterior Features	Likelihood of Malignancy	Pathology
1	0.520833	0.645833	0.229167	0.104167	0.541667	0.385459	0.770833
2	0.586957	0.826087	0.413043	0.239130	0.652174	0.291602	0.782609
3	0.723404	0.808511	0.297872	0.340426	0.574468	0.367778	0.744681
4	0.638298	0.808511	0.361702	0.297872	0.531915	0.342468	0.553191
5	0.652174	0.652174	0.326087	0.282609	0.500000	0.295426	0.717391
Average	0.624333	0.748223	0.325574	0.252841	0.560045	0.336546	0.713741

The average scores above show that the model does moderately well at predicting descriptors such as orientation and pathology, however, it performs poorly on descriptors like marginal zone and echo pattern. Upon intra-fold inspection, it appears that performance can vary greatly, upwards of $\pm 20\%$ across all descriptors regardless of their averaged performance. This phenomenon indicates that the results from the modified BI-RADS-Net are heavily dependent on which observations are placed into the training and validation sets, and it may suggest that the size of the current dataset is too small and too sensitive for building a robust model. Additionally, it was discussed in chapter 3 that the current Mayo Clinic dataset has a heavy imbalance among class distributions. This is particularly true in the descriptors that show a poor averaged performance in the table above, as some classes represent less than a few percentage points of the total distribution. While techniques were put in place to combat this issue, such as the application of class weights in the loss function and the relaxation of the mixing factor λ to favor minority classes, the dataset's imbalance may be too extreme in its current state to reliability train a robust multitask model.

Comparison to Trained Professional

Since the purpose of this study is to assess the use of deep learning to aid radiologists in the interpretation of BUS images, this section reviews the performance of a trained radiologist in predicting the pathological outcomes of a patient's study within the Mayo Clinic dataset. A major objective of this project is to build a model that can meet or exceed human-level performance thereby adding value in a clinical setting. By measuring a radiologist's performance, a baseline can be developed for evaluating model performance to understand if a model will meet this major objective.

Each of the 126 patient studies in the Mayo Clinic dataset contained a predicted histology assigned by the radiologist that indicated their professional opinion on whether the lesion was benign or malignant. A biopsy revealed the true pathology of the lesion, and a comparison is made to the predicted histology. Table 10 displays the diagnostic performance of the radiologist.

Table 10

Trained Radiologist Pathology Confusion Matrix

n=234		Predicted Values: Malignant	Predicted Values: Benign
Actual Values	Malignant	115	8
	Benign	78	33

Note. Recall that a majority of the 126 patients have both the transversal and longitudinal view, which accounts for the 234 images listed in the confusion matrix.

From the results above, it is clear that the radiologist does a very good job at identifying the truly malignant lesions as malignant (115) as well as avoiding misclassifying malignant lesions as benign (8); their measure of sensitivity is fine-tuned. Although the radiologist does a very good job at catching malignant lesions, it appears to come at the cost of potentially overclassifying benign lesions as

malignant (78), therefore reducing their overall measure of precision. Recall a lower precision generally equates to increases in the number of unnecessary medical procedures and increases in overall medical costs. Therefore, the aim is to provide a model that elevates both sensitivity *and* precision using the F1 score.

Now that the performance of the trained radiologist is analyzed, a comparison can be made between their performance and the ResNet-34-based modified BI-RADS-Net model selected from the previous section. Table 11 displays the same previously discussed performance metrics for both the radiologist and the ResNet-34 model.

Table 11

Comparison Between Radiologist and ResNet-34 Model

Method	F1 Score	Sensitivity	Precision	AUC	Accuracy
Radiologist	0.727848	0.934959	0.595854	0.616128	0.632478
ResNet-34	**0.743295**	0.788617	0.702898	0.709624	0.713675

As was eluded to above, the sensitivity of the radiologist is quite high compared to the ResNet-34 model. However, the lower precision of the radiologist ultimately produced a lower F1 score, therefore making the ResNet-34 model slightly more appealing in its ability to balance both sensitivity and precision. While the goal of this study was not to outperform a radiologist, seeing comparable results is a positive outcome that suggests that deep learning is a viable approach toward aiding radiologists with the task of BUS interpretation.

Summary

This chapter discussed the results and findings of the study using the methodologies described in chapter 3. The performance of a trained radiologist, along with three different architectural backbones that powered the modified BI-RADS-Net, were evaluated in their ability to predict BUS lesion

pathology. Additional performance metrics were calculated and used to make comparisons between the radiologist and the best model identified. F1 score was selected as the primary metric of comparison since it represents an overall balance between sensitivity and precision, both of which are important in a healthcare setting. The results showed that the ResNet-34 slightly outperformed the radiologist in its ability to predict pathology. However, it was noted that performance only differed by a few points, and the current Mayo Clinic dataset may be too small to make any conclusions on performance. Despite the challenges, these findings suggest that deep learning appears to be a viable approach to building a CAD system for the interpretation of BUS imaging at the Mayo Clinic.

Chapter 5: Discussion

Introduction

The purpose of this study focused on the application of deep learning models to standardize processes and aid radiologists in the clinical interpretation of BUS images. This final chapter will include a summary and a discussion of the research findings, make recommendations for future research, and provide a brief closing statement.

Summary of Findings

Using the entire dataset made available from the Mayo Clinic, training via five-fold cross-validation found that the ResNet-34-based modified BI-RADS-Net performed the best when compared to the other model architectures considered in this project. F1 score was selected as the comparison metric of choice since it balances sensitivity and precision, both of which are considered important measurements within the healthcare industry. Furthermore, ResNet-34 slightly outperformed a trained radiologist in the ability to predict whether a lesion was benign or malignant, however, the difference was minimal and may change as models are retrained when more data is made available.

An evaluation of the explainable multitask output of the ResNet-34 model discovered mixed performance toward the prediction of lesion characteristics. Descriptors such as orientation and pathology had a moderate level of performance, whereas descriptors like marginal zone and echo pattern performed poorly. Upon intra-fold inspection, all descriptors displayed a range of roughly $\pm 20\%$ in accuracy, regardless of their averaged outcome. This suggests the performance of the multitask model is very sensitive to the observations and distribution of classes within the training and validation sets.

Discussion

One major challenge with the real-world application of deep learning is directly correlated with the amount of data available to train and test the variety of models produced during a study. This is particularly true for medical imaging and the healthcare industry, where data is limited due to strict federal regulations and the sensitive nature of patient information. This challenge is further compounded by the need for a medical professional to correctly classify and annotate observations within the dataset for researchers to apply supervised learning.

Despite the limited data available within the Mayo Clinic dataset at the time of this writing, the best model created performed on par with a trained radiologist in its ability to predict lesion pathology. Domain-appropriate data augmentations such as the rotate, flip, zoom, and crop, played a key role in getting the most out of the limited dataset. Additionally, this project benefited particularly well from the mixup augmentation technique. By combining multiple images during training, the models needed to focus more intensely on learning features that uniquely describe the various class labels, as opposed to simply memorizing them. Although mixup required longer training sessions, it helped prevent the models from overfitting.

Model explainability is another important characteristic in the field of healthcare. As discussed in chapter 1, state-of-the-art deep learning models generally appear as black boxes since they contain many complex hidden layers and millions or billions of parameters. While all these parameters and complexities in architecture allow deep learning models to generalize well toward their task, it leaves their inner workings somewhat of a mystery. By including BUS lesion characteristics as model outputs, such as shape and orientation, radiologists receive a richer understanding of what went into the model's pathological prediction. Furthermore, a continuous value for the likelihood of malignancy was produced, providing radiologists with a better explanation of model output by reporting precisely where on the BI-RADS scale the model suggests. This approach distinctly targets the challenging boundary between BI-RADS category 3 and 4a where recommended management is either a plan of continued surveillance or a tissue biopsy.

Although this project successfully implemented the multitask approach described above, performance across the tasks was mixed. As described in the previous section, some lesion characteristics displayed moderate levels of performance, whereas others exhibited poor performance. This was due in part to the extremely imbalanced class distributions found within the current Mayo Clinic dataset. Some classes only represent a percentage point or two of the total class distribution within their respective descriptor. Although techniques were implemented to combat the imbalanced dataset, such as class weights and remix, they did not have enough of an impact to reach the desired levels of performance across all descriptors. Future releases of the dataset should focus on minority class examples and multitask performance should be reassessed. Additional techniques such as saliency maps (Shen et al., 2021) or layer-wise relevance propagation (Böhle et al., 2019) could also be explored as alternative methods for meeting explainability requirements. As such, saliency maps are described in the next section on future research.

In summary, this project showed that deep learning is a powerful approach toward building a CAD system that can aid radiologists in the task of BUS interpretation at the Mayo Clinic. While some objectives did not perform as well as expected, the primary cause appears to be caused from a lack of available data. Since the model architectures and training techniques described within this research paper are not reliant upon the specific dataset, they can easily be applied again once more data is made available from Mayo Clinic.

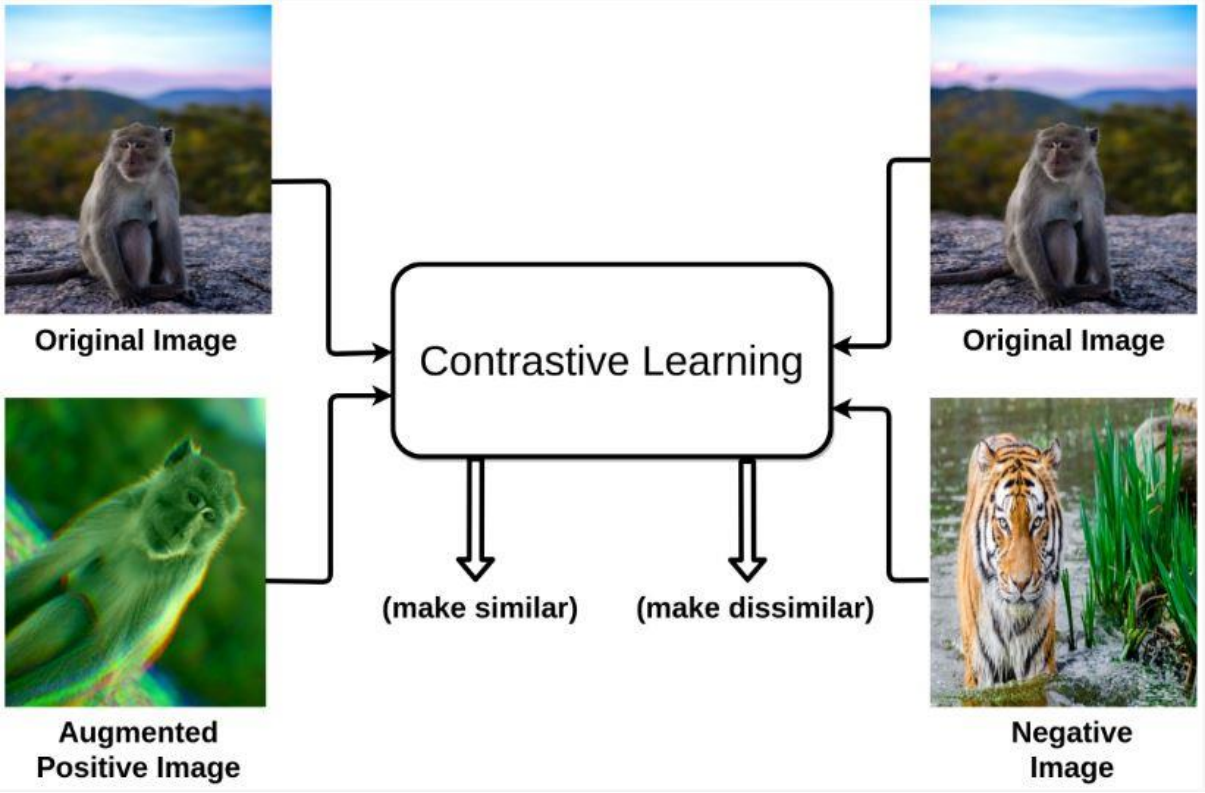
Suggestions for Future Research

This research has made a significant contribution to the application of deep learning in the interpretation of BUS imaging at the Mayo Clinic. However, as was described in the previous section, the methods proposed within this paper still leave room for improvement. First, an approach to realistically learn from a much larger dataset is explored. Second, since it is difficult to predict exactly how the multitask models studied within this project will perform once retrained on additional data, an additional explainability technique is considered.

It should be clear by now that limited amounts of data pose a significant challenge to building very data-hungry deep learning models that can generalize well at scale. What makes the medical imaging domain particularly challenging is finding a trained resource to label and annotate raw data when it becomes available; opportunities to employ practicing radiologists are rare. A recent technique known as self-supervised learning aims to solve the challenge of human-driven labeling by doing away with labels entirely. This technique uses parts of its own image, or augmentations of itself, to learn about latent features embedded within that make it unique from other images within a dataset (Jaiswal et al., 2020). Figure 11 displays the idea behind contrastive learning, which is a form of self-supervised learning.

Figure 11

Contrastive Learning – a Form of Self-Supervised Learning (Jaiswal et al., 2020)

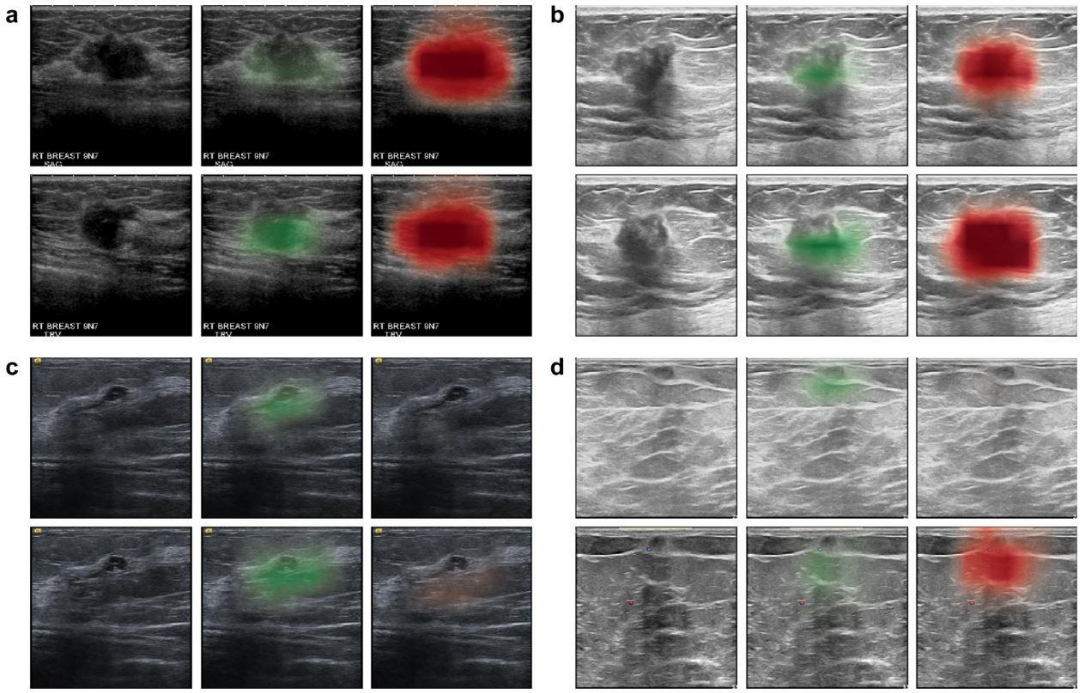


Once a self-supervised learning model completes its pretext task of learning to separate images into their classes, its parameters can transfer toward a downstream target task (Noroozi et al., 2018). At the time of this writing, project stakeholders at the Mayo Clinic and the University of Wisconsin – La Crosse were in a unique position to overcome one of the main challenges of the medical imaging domain, data availability. Efforts were underway to automate the de-identification and delivery of vast quantities of high-quality BUS images from the Mayo Clinic. However, much of this data would be unlabeled upon its initial release and may not be realistic to annotate in its entirety, therefore making self-supervised learning a promising potential solution. Parameters learned from this vast quantity of unlabeled data have the potential to be transferred into the supervised learning architectures and training processes explored within this research.

The second area of continued research is to explore additional explainability techniques. One proposed technique is known as a saliency map, which highlights unique features of the images that correlate to the potential outcomes predicted by the model. Figure 12 displays an example of saliency maps applied to BUS imaging.

Figure 12

The Application of Saliency Maps on a Sample of BUS Images (Shen et al., 2021)



In the above images, each study (A-D) consists of the original BUS image (left) from both the longitudinal and transversal views. Features indicative of a benign pathology are displayed in the center using a green shading, whereas images on the right indicate malignant features. It is clear to see studies A and B have a much higher suggestion for malignancy, whereas study C appears to be more benign. Study D appears suggestive of malignancy. Shen et al. (2021) found their proposed saliency maps, although not perfect, outperformed the majority of a panel study of 10 radiologists. The ability to visualize predictive outputs via the implementation of saliency maps is just one of many other ways to improve model explainability and instill confidence and trust into users of the system.

Conclusion

This study emphasized that current state-of-the-art deep learning models, with the use of proper techniques and enough data, have the potential to aid radiologists in their interpretation of BUS images. While many historical studies have shown promise in the past, they were considered untenable due to the explainability requirements needed from the healthcare industry. In recent years, techniques to address both the challenges of explainability and data scarcity have evolved to a point where it becomes entirely feasible to develop a functional CAD system in the clinical setting. As more data is made available and additional techniques implemented, model performance and the value add is only likely to continue increasing. We look forward to progressing into future phases of this project to accomplish the original stakeholder objectives of reducing unnecessary medical procedures and costs, and above all else, delivering a consistent and exemplary patient experience.

References

Abdullah, N., Mesurolle, B., El-Khoury, M., & Kao, E. (2009). Breast Imaging Reporting and Data System Lexicon for US: Interobserver agreement for assessment of Breast Masses. *Radiology, 252*(3), 665–672. https://doi.org/10.1148/radiol.2523080670

Alto, V. (2020). *Data augmentation in Deep Learning.* Medium. Retrieved June 25, 2022, from https://medium.com/analytics-vidhya/data-augmentation-in-deep-learning-3d7a539f7a28

American Cancer Society. (2022). *Breast cancer statistics: How common is breast cancer?* Retrieved June 11, 2022, from https://www.cancer.org/cancer/breast-cancer/about/how-common-is-breast-cancer.html

American College of Radiology. (2017). *Breast density brochure-nov2017 F*. Retrieved June 12, 2022, from https://www.acr.org/-/media/ACR/Files/Breast-Imaging-Resources/Breast-Density-bro_ACR_SBI.pdf

American College of Radiology. (n.d.). *Breast Imaging Reporting & Data System*. Retrieved June 14, 2022, from https://www.acr.org/Clinical-Resources/Reporting-and-Data-Systems/Bi-Rads

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). https://doi.org/10.1093/jnci/djv367

Berry, D. A., Cronin, K. A., Plevritis, S. K., Fryback, D. G., Clarke, L., Zelen, M., Mandelblatt, J. S., Yakovlev, A. Y., Habbema, J. D., & Feuer, E. J. (2005). Effect of screening and adjuvant therapy on mortality from breast cancer. *New England Journal of Medicine, 353*(17), 1784–1792. https://doi.org/10.1056/nejmoa050518

Birnbaum, J. K., Duggan, C., Anderson, B. O., & Etzioni, R. (2018). Early detection and treatment strategies for breast cancer in low-income and upper middle-income countries: A modelling study. *The Lancet Global Health, 6*(8). https://doi.org/10.1016/s2214-109x(18)30257-2

Böhle, M., Eitel, F., Weygandt, M., & Ritter, K. (2019). Layer-wise relevance propagation for explaining deep neural network decisions in MRI-based alzheimer's disease classification. *Frontiers in Aging Neuroscience, 11*. https://doi.org/10.3389/fnagi.2019.00194

Boyd, N. F., Guo, H., Martin, L. J., Sun, L., Stone, J., Fishell, E., Jong, R. A., Hislop, G., Chiarelli, A., Minkin, S., & Yaffe, M. J. (2007). Mammographic density and the risk and detection of breast cancer. *New England Journal of Medicine, 356*(3), 227–236. https://doi.org/10.1056/nejmoa062790

Brownlee, J. (2021). *Gentle introduction to the Adam optimization algorithm for deep learning*. Machine Learning Mastery. Retrieved July 11, 2022, from https://machinelearningmastery.com/adam-optimization-algorithm-for-deep-learning/

Burnside, E. S., Sickles, E. A., Bassett, L. W., Rubin, D. L., Lee, C. H., Ikeda, D. M., Mendelson, E. B., Wilcox, P. A., Butler, P. F., & D'Orsi, C. J. (2009). The ACR BI-RADS® experience: Learning from history. *Journal of the American College of Radiology, 6*(12), 851–860. https://doi.org/10.1016/j.jacr.2009.07.023

Byra, M. (2021). Breast mass classification with transfer learning based on scaling of deep representations. *Biomedical Signal Processing and Control, 69*, 102828. https://doi.org/10.1016/j.bspc.2021.102828

Caruana, R. (1997). *Machine Learning, 28*(1), 41–75. https://doi.org/10.1023/a:1007379606734

Chae, E. Y., Kim, H. H., Cha, J. H., Shin, H. J., & Kim, H. (2013). Evaluation of screening whole-breast sonography as a supplemental tool in conjunction with mammography in women with dense

breasts. *Journal of Ultrasound in Medicine, 32*(9), 1573–1578. https://doi.org/10.7863/ultra.32.9.1573

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16*, 321–357. https://doi.org/10.1613/jair.953

Chen, D.-R., & Hsiao, Y.-H. (2008). Computer-aided diagnosis in breast ultrasound. *Journal of Medical Ultrasound, 16*(1), 46–56. https://doi.org/10.1016/s0929-6441(08)60005-3

Chen, Z., Wang, X., Xie, X., Wu, T., Bu, G., Wang, Y., & Chen, E. (2019). Co-attentive multi-task learning for explainable recommendation. *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence.* https://doi.org/10.24963/ijcai.2019/296

Chou, H.-P., Chang, S.-C., Pan, J.-Y., Wei, W., & Juan, D.-C. (2020). *Remix: Rebalanced mixup.* arXiv.org. Retrieved July 5, 2022, from https://doi.org/10.48550/arXiv.2007.03943

DeSantis, C. E., Bray, F., Ferlay, J., Lortet-Tieulent, J., Anderson, B. O., & Jemal, A. (2015). International variation in female breast cancer incidence and mortality rates. *Cancer Epidemiology, Biomarkers & Prevention, 24*(10), 1495–1506. https://doi.org/10.1158/1055-9965.epi-15-0535

Dense Breasts Canada. (2022). Breast density matters. Retrieved June 12, 2022, from https://densebreastscanada.ca/

Ding, J., Cheng, H. D., Huang, J., Liu, J., & Zhang, Y. (2012). Breast ultrasound image classification based on multiple-instance learning. *Journal of Digital Imaging, 25(*5), 620–627. https://doi.org/10.1007/s10278-012-9499-x

Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., & Darrell, T. (2013). *DeCAF: A deep convolutional activation feature for generic visual recognition*. arXiv.org. Retrieved June 25, 2022, from https://doi.org/10.48550/arXiv.1310.1531

Ezzat, D., Afify, H. M., Taha, M. H., & Hassanien, A. E. (2020). Convolutional neural network with batch normalization for classification of endoscopic gastrointestinal diseases. *Studies in Big Data*, 113–128. https://doi.org/10.1007/978-3-030-59338-4_7

Fumo, D. (2018). *Why is everyone talking about Artificial Intelligence?* Medium. Retrieved June 16, 2022, from https://towardsdatascience.com/why-is-everyone-talking-about-ai-73bab31bf9c1

Galdran, A., Carneiro, G., & Ballester, M. A. G. (2021). *Balanced-mixup for highly imbalanced medical image classification*. arXiv.org. Retrieved June 26, 2022, from https://doi.org/10.48550/arXiv.2109.09850

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning.* MIT Press. http://www.deeplearningbook.org

Gulum, M. A., Trombley, C. M., & Kantardzic, M. (2021). A review of Explainable Deep Learning Cancer Detection Models in medical imaging. *Applied Sciences, 11*(10), 4573. https://doi.org/10.3390/app11104573

He, K., Zhang, X., Ren, S., & Sun, J. (2015). *Deep residual learning for image recognition*. arXiv.org. Retrieved July 20, 2022, from https://doi.org/10.48550/arXiv.1512.03385

Huang, G., Liu, Z., van der Maaten, L., & Weinberger, K. Q. (2018). *Densely connected Convolutional Networks*. arXiv.org. Retrieved July 22, 2022, from https://doi.org/10.48550/arXiv.1608.06993

Huh, M., Agrawal, P., & Efros, A. A. (2016). *What makes ImageNet good for transfer learning?* arXiv.org. Retrieved June 25, 2022, from https://doi.org/10.48550/arXiv.1608.08614

Hussain, Z., Gimenez, F., Yi, D., & Rubin, D. (2018). Differential Data Augmentation Techniques for Medical Imaging Classification Tasks. *Annual Symposium proceedings. AMIA Symposium, 2017,* 979–984. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5977656/

IBM. (n.d.). *What is overfitting?* IBM. Retrieved June 26, 2022, from https://www.ibm.com/cloud/learn/overfitting

Ioffe, S., & Szegedy, C. (2015). *Batch normalization: Accelerating deep network training by reducing internal covariate shift*. arXiv.org. Retrieved June 26, 2022, from https://arxiv.org/abs/1502.03167

Jaiswal, A., Babu, A. R., Zadeh, M. Z., Banerjee, D., & Makedon, F. (2020). A survey on contrastive self-supervised learning. *Technologies, 9*(1), 2. https://doi.org/10.3390/technologies9010002

Jakubowski, W., Dobruch-Sobczak, K., & Migda, B. (2012). Errors and mistakes in breast ultrasound diagnostics. *Journal of Ultrasonography, 12*(50), 286–298. https://doi.org/10.15557/jou.2012.0014

Joy, J. E., Penhoet, E. E., & Petitti D. B. (2005). *Saving women's lives: Strategies for improving breast cancer detection and diagnosis.* The National Academies Press. https://www.ncbi.nlm.nih.gov/books/NBK22311/

Kaplan, S. S. (2001). Clinical utility of bilateral whole-breast us in the evaluation of women with dense breast tissue. *Radiology, 221*(3), 641–649. https://doi.org/10.1148/radiol.2213010364

Khatua, D. P. (2020). *Easy way to improve image classifier performance(part 1) - mixup augmentation with codes*. Medium. Retrieved June 26, 2022, from https://medium.com/@wolframalphav1.0/easy-way-to-improve-image-classifier-performance-part-1-mixup-augmentation-with-codes-33288db92de5

Kingma, D. P., Ba, J. (2017). *Adam: A method for stochastic optimization*. arXiv.org. Retrieved July 9, 2022, from https://doi.org/10.48550/arXiv.1412.6980

Kohli, M. D., Summers, R. M., & Geis, J. R. (2017). *Medical Image Data and datasets in the era of machine learning-whitepaper from the 2016 C-Mimi Meeting Dataset Session - Journal of Digital Imaging*. SpringerLink. Retrieved June 25, 2022, from https://doi.org/10.1007/s10278-017-9976-3

Lazarus, E., Mainiero, M. B., Schepps, B., Koelliker, S. L., & Livingston, L. S. (2006). BI-RADS Lexicon for US and mammography: Interobserver variability and positive predictive value. *Radiology, 239*(2), 385–391. https://doi.org/10.1148/radiol.2392042127

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Computation, 1*(4), 541–551. https://doi.org/10.1162/neco.1989.1.4.541

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE, 86*(11), 2278–2324. https://doi.org/10.1109/5.726791

Marques, S., Schiavo, F., Ferreira, C. A., Pedrosa, J., Cunha, A., & Campilho, A. (2021). A multi-task CNN approach for lung nodule malignancy classification and characterization. *Expert Systems with Applications, 184*, 115469. https://doi.org/10.1016/j.eswa.2021.115469

Morid, M. A., Borjali, A., & Del Fiol, G. (2021). A scoping review of transfer learning research on medical image analysis using ImageNet. *Computers in Biology and Medicine, 128*, 104115. https://doi.org/10.1016/j.compbiomed.2020.104115

National Cancer Institute. (n.d.). *Mammograms*. Retrieved June 20, 2022, from https://www.cancer.gov/types/breast/mammograms-fact-sheet#what-are-the-benefits-and-potential-harms-of-screening-mammograms

Noroozi, M., Vinjimoor, A., Favaro, P., & Pirsiavash, H. (2018). *Boosting self-supervised learning via knowledge transfer.* Retrieved July 15, 2022, from https://openaccess.thecvf.com/content_cvpr_2018/papers/Noroozi_Boosting_Self-Supervised_Learning_CVPR_2018_paper.pdf

Peconic Bay Medical Center. (n.d.). *Latest advancements in ultrasound imaging technology*. Retrieved June 15, 2022, from https://www.pbmchealth.org/news-events/blog/latest-advancements-ultrasound-imaging-technology

Pham, H., Dai, Z., Xie, Q., Luong, M-T., & Le, Q. V. (2021). *Meta pseudo labels*. arXiv.org. Retrieved June 20, 2022, from https://doi.org/10.48550/arXiv.2003.10580

Razavian, A. S., Azizpour, H., Sullivan, J., & Carlsson, S. (2014). *CNN features off-the-shelf: An astounding baseline for recognition*. arXiv.org. Retrieved June 25, 2022, from https://doi.org/10.48550/arXiv.1403.6382

Romero, A. (2021). *GPT-3 - A complete overview*. Medium. Retrieved June 17, 2022, from https://towardsdatascience.com/gpt-3-a-complete-overview-190232eb25fd

Roth, H. R., Farag, A., Lu, L., Turkbey, E. B., & Summers, R. M. (2015). Deep convolutional networks for pancreas segmentation in CT imaging. *SPIE Proceedings*. https://doi.org/10.1117/12.2081420

Ruder, S. (2017). *An overview of multi-task learning in Deep Neural Networks*. arXiv.org. Retrieved June 27, 2022, from https://doi.org/10.48550/arXiv.1706.05098

Saha, S. (2018). *A comprehensive guide to Convolutional Neural Networks - the eli5 way*. Medium. Retrieved June 22, 2022, from https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53

Sarvamangala, D. R., & Kulkarni, R. V. (2021). Convolutional neural networks in medical image understanding: A survey. *Evolutionary Intelligence, 15*(1), 1–22. https://doi.org/10.1007/s12065-020-00540-3

Shen, D., Wu, G., & Suk, H.-I. (2017). Deep learning in medical image analysis. *Annual Review of Biomedical Engineering, 19*(1), 221–248. https://doi.org/10.1146/annurev-bioeng-071516-044442

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

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for Deep Learning. *Journal of Big Data, 6*(1). https://doi.org/10.1186/s40537-019-0197-0

Simonyan, K., & Zisserman, A. (2015). *Very deep convolutional networks for large-scale image recognition*. arXiv.org. Retrieved July 20, 2022, from https://doi.org/10.48550/arXiv.1409.1556

Singh, A., Sengupta, S., & Lakshminarayanan, V. (2020). *Explainable deep learning models in medical image analysis*. arXiv.org. Retrieved June 17, 2022, from https://doi.org/10.48550/arXiv.2005.13799

Smith, L. N. (2017). *Cyclical learning rates for training neural networks*. arXiv.org. Retrieved July 11, 2022, from https://doi.org/10.48550/arXiv.1506.01186

Sowjanya, A. M., & Mrudula, O. (2022). Effective treatment of imbalanced datasets in health care using modified smote coupled with stacked deep learning algorithms. *Applied Nanoscience*. https://doi.org/10.1007/s13204-021-02063-4

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research, 15*(56), 1929-1958. https://jmlr.org/papers/v15/srivastava14a.html

Sun, C., Shrivastava, A., Singh, S., & Gupta, A. (2017). *Revisiting unreasonable effectiveness of data in deep learning era*. arXiv.org. Retrieved June 24, 2022, from https://doi.org/10.48550/arXiv.1707.02968

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2015). *Rethinking the inception architecture for computer vision*. arXiv.org. Retrieved June 20, 2022, from https://doi.org/10.48550/arXiv.1512.00567

Thigpen, D., Kappler, A., & Brem, R. (2018). The role of ultrasound in screening dense breasts—a review of the literature and practical solutions for implementation. *Diagnostics, 8*(1), 20. https://doi.org/10.3390/diagnostics8010020

Vignesh, S. (2020). *The world through the eyes of CNN*. Medium. Retrieved June 25, 2022, from https://medium.com/analytics-vidhya/the-world-through-the-eyes-of-cnn-5a52c034dbeb

Wang, C., Liu, Y., Wang, F., Zhang, C., Wang, Y., Yuan, M., & Yang, G. (2022). *Towards reliable and explainable AI model for solid pulmonary nodule diagnosis.* arXiv.org. Retrieved June 27, 2022, from https://doi.org/10.48550/arXiv.2204.04219

Wang, D., Zhang, Y., Zhang, K., & Wang, L. (2020). *FOCALMIX: Semi-supervised learning for 3D Medical Image Detection*. arXiv.org. Retrieved June 26, 2022, from https://doi.org/10.48550/arXiv.2003.09108

Weiss, K., Khoshgoftaar, T. M., & Wang, D. D. (2016). A survey of Transfer Learning. *Journal of Big Data, 3*(1). https://doi.org/10.1186/s40537-016-0043-6

World Health Organization. (2021). *Breast cancer.* Retrieved June 11, 2022, from https://www.who.int/news-room/fact-sheets/detail/breast-cancer

Xie, J., Xu, L., & Chen, E. (2012). Image denoising and inpainting with Deep Neural Networks. *Advances in Neural Information Processing Systems.* Retrieved July 2, 2022, from https://proceedings.neurips.cc/paper/2012/hash/6cdd60ea0045eb7a6ec44c54d29ed402-Abstract.html

Zhang, B., Vakanski, A., & Xian, M. (2021). Bi-Rads-Net: An explainable multitask learning approach for cancer diagnosis in breast ultrasound images. *2021 IEEE 31st International Workshop on Machine Learning for Signal Processing (MLSP).* https://doi.org/10.1109/mlsp52302.2021.9596314

Zhang, H., Cisse, M., Dauphin, Y. N., & Lopez-Paz, D. (2018). *Mixup: Beyond empirical risk minimization.* arXiv.org. Retrieved June 26, 2022, from https://arxiv.org/abs/1710.09412

Zhang, Q., Xiao, Y., Dai, W., Suo, J., Wang, C., Shi, J., & Zheng, H. (2016). Deep learning based classification of breast tumors with shear-wave elastography. *Ultrasonics, 72*, 150–157. https://doi.org/10.1016/j.ultras.2016.08.004

Appendix A: Code

The code for this project was created using Python and Juypter notebooks, running on Google Collaboratory for access to GPUs. The notebook is located at: https://github.com/joshjarvey/DS785-BUS-Interpretation