Classification of Breast Lesions using Deep Learning

Justin Hall

University of Wisconsin – La Crosse

DS: 785 Capstone

12/12/2021

Abstract

Breast ultrasounds are a standard modality used to find cancerous lesions and provide early care for patients. However, this modality is sensitive to user dependency. The discrepancies between radiologists' readings for the same image may lead to inadequate patient care and inconsistent outcomes. As breast cancer continues to affect people around the globe, a new approach is needed to detect cancerous lesions. To provide a system for standardized classifications of these images, we examine the use of state-of-the-art convolutional neural network architectures to improve patient care. Our analysis showed that while all models showed similar AUC on our validation dataset of 121 images, DenseNet-201 maximized the F1-score. To provide confidence in our predictions, we then use local interpretable model-agnostic explanations (LIME) to find what image regions are essential to our model. Our model identified areas of lesion consistent with areas identified by trained radiologists suggesting that this model shows promise in aiding radiologists resulting in better care for patients.

Acknowledgments

I want to thank Mayo Clinic Health System and Dr. Song Chen, and Dr. Jeff Baggett from the University of Wisconsin – La Crosse for allowing me to work on this project.

Secondly, I would like to thank Dr. Rich Ellis for his willingness to offer guidance and share his knowledge in breast imaging. Without his work annotating and classifying images, this project would not be possible.

I would also like to thank the other researchers on the project: Adam Silberfein, David Halama, Simon Wagner, Suriya Mohan, and Lucas Spellman. Our discussions and collaboration helped challenge and inspire me to try new techniques and grow my knowledge of machine learning.

Lastly, I would like to thank my partner Jennifer and our daughter Gwendolyn. Their constant support and encouragement helped me get through my studies.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Background

The Mayo Clinic Health System provides many services, including mammography and breast ultrasounds imaging for Wisconsin, Minnesota, and Iowa. They are committed to providing quality, affordable, and specialized care to every patient. This study is part of a collaboration between Mayo Clinic Health System and the University of Wisconsin – La Crosse to develop a state-of-the-art system to aid in the interpretation of breast ultrasound (BUS) images.

According to a recent study, breast cancer now affects more people globally than any other form of cancer (Breast cancer now most common form of cancer: WHO taking action, 2021). It is estimated that over 280,000 new cases will be diagnosed in the United States alone in 2021, resulting in over 43,000 deaths (U.S. Breast Cancer Statistics, 2021). A critical factor in reducing the number of mortalities is ensuring that potentially life-threatening lesions are found early (Sun et al., 2017). By finding these lesions early, patients have access to more treatment options when they are most effective.

To aid in early detection, doctors recommend regular mammograms or physical examinations to help identify abnormalities in breast tissue (Breast Cancer Early Detection and Diagnosis, 2021). Another common alternative is the use of ultrasound imaging. Breast ultrasounds offer a non-radioactive imaging technique, allowing for a safe and non-invasive way to assess any conspicuous masses and limiting the need for unnecessary surgery or invasive procedures such as a biopsy.

BUS Interpretation

The Breast Imaging Reporting and Data System (BI-RADS) provide a standardized way for radiologists to describe breast lesions found in BUS images. This standardization aims to improve the quality of the assessment, improving patient care. The system defines seven levels of ranking 0 – 6, where a higher score reflects an increased likelihood of malignancy. The probability of malignancy for each score is shown in Table 1 below (Mendelson et al., 2013):

Table 1

BI-RADS Score and Probability of Malignancy

Score	Classification	Probability of Malignancy
0	Incomplete	N/A
1	Negative	0%
2	Benign	0%
3	Probably Benign	<2%
4	Suspicious for Malignancy	2-94%
5	Highly Suggestive of Malignancy	>95%
6	Known Malignancy	100%

Statement of Problem

Advancements in medical imaging have greatly improved the level of diagnostics that can be recorded. These advancements allow radiologists to get a detailed view of breast tissues and vascularity, allowing for more accurate diagnoses and better patient care. Despite these developments, the interpretations of breast ultrasounds still rely heavily on the experience and judgment of radiologists to define characteristics found in the images. Many features described in the BI-RAD assessment can be found in benign and malignant lesions. The subjectivity of these characteristics can lead to wildly varying assessments across radiologists.

Because BUS images are often a determining factor for the need for a biopsy, low sensitivity in identifying malignant lesions results in unnecessary procedures—the user-reliant classification of BUS imaging results in non-uniform patient care. According to internal studies at Mayo Clinic Health System, the percentage of positive breast ultrasound biopsies varies as much as 51%. For many patients, this means unnecessary invasive procedures and increased medical costs.

Purpose of the Study

With the recent improvements in computer vision over the last decade, many industries have begun using machine learning models to aid in various tasks. The medical sector is one of these industries. Many academic studies have shown the success of deep learning algorithms to assist in classifying and detecting disease in medical images. Still, these algorithms have not done well to generalize to a clinical setting. This study aims to create a state-of-the-art computer-aided diagnosis (CAD) system to create a more standardized approach to assessing BUS images and assigning a diagnosis. Three methods will be developed to generate the CAD system:

1. Automated CAD using deep learning: This approach will apply cutting-edge models in computer vision such as state-of-the-art convolutional neural networks (CNNs), vision transformers, and self-supervised learning to segment and classify lesions in BUS images.
2. Automated CAD by mimicking human experts: Using characteristics of breast lesions that have been determined to be significant by human experts, a model will be built to

classify a lesion. A trained radiologist can input these characteristics or extract them from the image via image processing techniques.

3. Automated CAD using an ensemble of approaches: This approach will create an ensemble model using the models developed in methods one and two. This ensemble will hopefully provide an improved estimate of the probability of malignancy.

These three approaches were chosen to offer a mixture of accuracy and interpretability. Deep learning models are very good at discovering complex relationships and have proven highly accurate in many scenarios. However, these models typically are hard to decipher and operate as "black boxes." Method two offers a more human understanding of what features are essential to malignancy. The blend of techniques will give Mayo Clinic Health System a unique system that incorporates both human-injected features and new features found by deep learning.

Significance of the Study

Many existing studies have been performed on publicly available BUS imaging datasets. These studies have taken various approaches, including many different machine learning algorithms, showing great success on these datasets. However, very few of these systems have been implemented in a clinical setting. This CAD system will be unique in its blend of deep learning and expert human knowledge trained for a specific patient population. This system, if successful, could help Mayo Clinic Health System to better patient care and reduce medical costs system-wide.

Project Outline

This project will consist of three planned phases to deliver Mayo Clinic Health System updates on time. Each phase will work to provide a significant update to the CAD system and communicate progress and challenges. These three phases are described below:

- Phase one will consist of initial model building. The critical components of this phase will be establishing the best model architectures, defining essential data augmentation techniques, and hyperparameter tuning of these models.
- Phase two will work to make the models scalable. Additional patient information will also be added to the models, such as patient demographics and pre-test likelihood of cancer.
- Phase three will develop an application that radiologists can utilize in their workflows to produce real-time assessments using the models from the earlier phases.

Limitations

Researchers have been divided into two teams given the enormous scope of the CAD system defined above. Each team will be responsible for one of the first two methods described above. The following chapters will focus on my primary roles in developing models, primarily the use of CNNs, for classification using method one of phase one of the project.

Chapter 2: Literature Review

Though commonly an adjunctive modality to mammography, breast ultrasound offers a safe, painless, and effective classification of potentially cancerous areas. Still, the user-dependent disadvantages to ultrasound images have made them a common target for computer-aided diagnosis (CAD) systems. These systems have taken various approaches, including clustering algorithms, support vector machines, and decision tree classifiers.

However, recent advancements in deep learning have become state-of-the-art for nearly all computer vision tasks. This literature review will focus on the applications of supervised deep learning algorithms, primarily those utilizing convolutional neural networks (CNNs), to aid in diagnosing BUS images.

Deep Learning Architecture

Improvements in imaging devices have substantially increased the quality of medical images, resulting in large high-resolution data. Despite these advancements, interpretation still relies on the experience of trained individuals who are prone to human error. Traditional machine learning techniques have been applied to aid radiologists with some success (Gómez Flores et al., 2015; Vijayarajeswari et al., 2019). However, these techniques struggle to extract the complex features associated with medical diagnoses (Razzak et al., n.d).

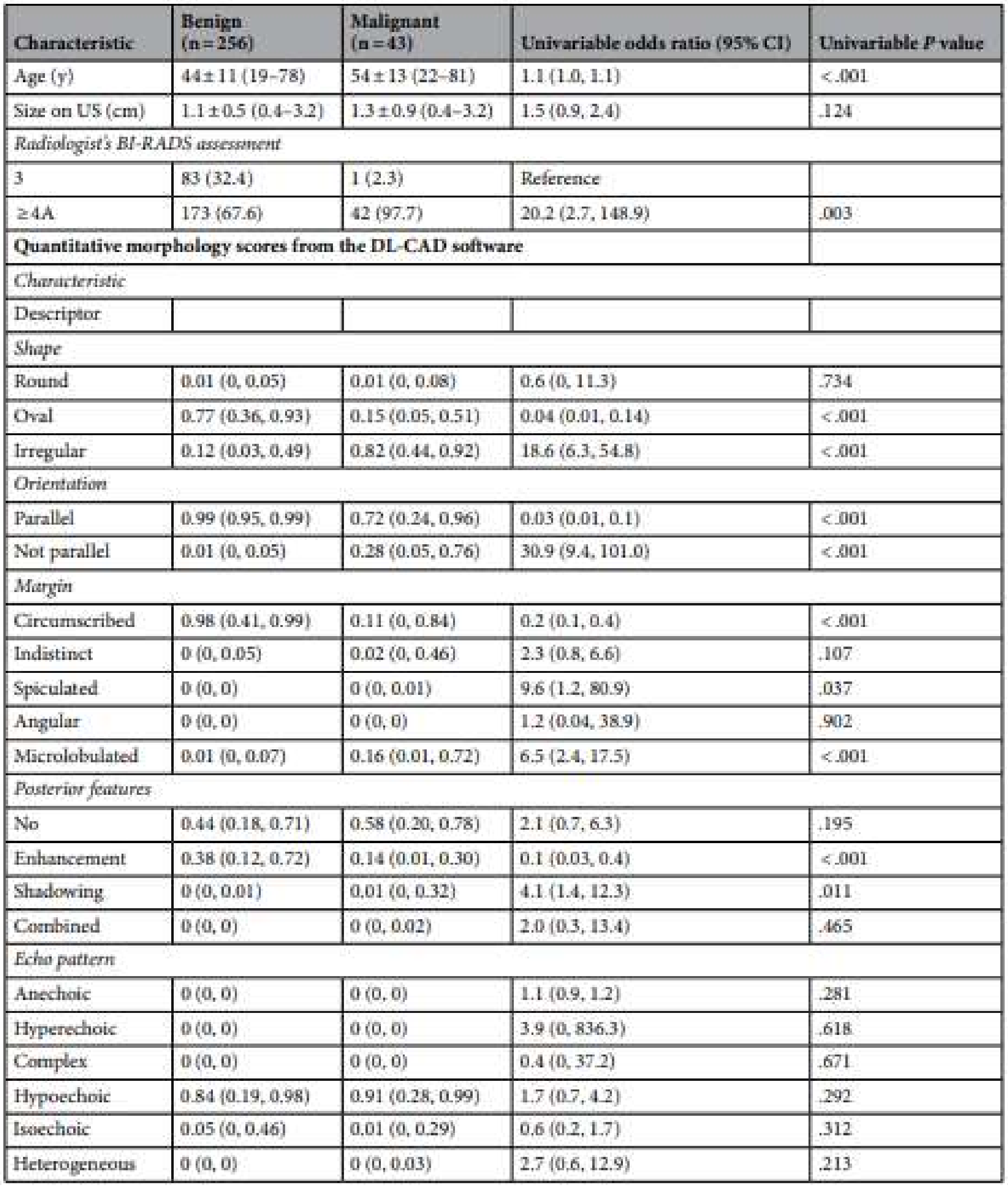
Deep learning models using CNNs are designed using a hierarchical structure. Each layer in the hierarchy can detect more complex features the deeper the layer is in the hierarchy. Unlike other machine learning techniques, CNNs do not rely on information gained from human experience or feature engineering to classify images. The features detected in each CNN layer are learned from the raw input. This means they can detect the obscure features known to domain experts and be used to detect new subtle and more intricate features not perceivable by humans. This ability has inspired the use of CNNs for many CAD systems.

Classification

As discussed in the above chapter, the Breast Imaging Reporting and Data System (BI-RADS) was developed to help radiologists classify lesions based on defined characteristics. Kim et al. (2021) analyzed the characteristics making up the cohort to determine which features are most important in determining malignancy. Their findings, displayed in Figure 1, show that only a small number of these categories were statistically different between benign and malignant lesions using a logistic regression model. This lends further evidence to suggest that the BI-RAD score with low sensitivity is not a good target for the classification task. As lesions are ultimately benign or malignant, most deep learning models are designed as binary classifiers (Byra, 2021; Saxena, 2021; Tanaka et al., 2019).

Figure 1

Statistical Importance of BI-RADS variables



Characteristic	Benign (n = 256)	Malignant (n = 43)	Univariable odds ratio (95% CI)	Univariable *P* value
Age (y)	44± 11 (19–78)	54 ± 13 (22–81)	1.1 (1.0, 1.1)	< .001
Size on US (cm)	1.1 ± 0.5 (0.4–3.2)	1.3 ± 0.9 (0.4–3.2)	1.5 (0.9, 2.4)	.124
Radiologist's BI-RADS assessment				
3	83 (32.4)	1 (2.3)	Reference	
≥4A	173 (67.6)	42 (97.7)	20.2 (2.7, 148.9)	.003
Quantitative morphology scores from the DL-CAD software				
Characteristic				
Descriptor				
Shape				
Round	0.01 (0, 0.05)	0.01 (0, 0.08)	0.6 (0, 11.3)	.734
Oval	0.77 (0.36, 0.93)	0.15 (0.05, 0.51)	0.04 (0.01, 0.14)	< .001
Irregular	0.12 (0.03, 0.49)	0.82 (0.44, 0.92)	18.6 (6.3, 54.8)	< .001
Orientation				
Parallel	0.99 (0.95, 0.99)	0.72 (0.24, 0.96)	0.03 (0.01, 0.1)	< .001
Not parallel	0.01 (0, 0.05)	0.28 (0.05, 0.76)	30.9 (9.4, 101.0)	< .001
Margin				
Circumscribed	0.98 (0.41, 0.99)	0.11 (0, 0.84)	0.2 (0.1, 0.4)	< .001
Indistinct	0 (0, 0.05)	0.02 (0, 0.46)	2.3 (0.8, 6.6)	.107
Spiculated	0 (0, 0)	0 (0, 0.01)	9.6 (1.2, 80.9)	.037
Angular	0 (0, 0)	0 (0, 0)	1.2 (0.04, 38.9)	.902
Microlobulated	0.01 (0, 0.07)	0.16 (0.01, 0.72)	6.5 (2.4, 17.5)	< .001
Posterior features				
No	0.44 (0.18, 0.71)	0.58 (0.20, 0.78)	2.1 (0.7, 6.3)	.195
Enhancement	0.38 (0.12, 0.72)	0.14 (0.01, 0.30)	0.1 (0.03, 0.4)	< .001
Shadowing	0 (0, 0.01)	0.01 (0, 0.32)	4.1 (1.4, 12.3)	.011
Combined	0 (0, 0)	0 (0, 0.02)	2.0 (0.3, 13.4)	.465
Echo pattern				
Anechoic	0 (0, 0)	0 (0, 0)	1.1 (0.9, 1.2)	.281
Hyperechoic	0 (0, 0)	0 (0, 0)	3.9 (0, 836.3)	.618
Complex	0 (0, 0)	0 (0, 0)	0.4 (0, 37.2)	.671
Hypoechoic	0.84 (0.19, 0.98)	0.91 (0.28, 0.99)	1.7 (0.7, 4.2)	.292
Isoechoic	0.05 (0, 0.46)	0.01 (0, 0.29)	0.6 (0.2, 1.7)	.312
Heterogeneous	0 (0, 0)	0 (0, 0.03)	2.7 (0.6, 12.9)	.213

Techniques for improving performance

The performance of these deep learning models depends many times on the volume of data that can be used for training (Lee et al., 2017). However, large training sets like ImageNet, a popular dataset for testing, took years to assemble (Deng et al., 2010). Because medical images must be labeled by trained radiologists and are subject to ethical and legal issues, assembling large datasets is often difficult. Many of the datasets found in the literature consist of at most a few thousand images (Byra et al., 2021; Saxena et al., 2021; Chiao et al., 2021; Huynh et al., 2016). Much of the literature utilized two different techniques for alleviating this issue.

Transfer Learning

To mitigate the limitation of small datasets in medical imaging, researchers often use transfer learning. Transfer learning allows neural networks, particularly CNNs, to be versatile by utilizing pre-trained networks to function as starting points for new tasks. Furthermore, transfer learning has been shown to significantly decrease the training times of CNNs and increase performance compared to models trained with random initial parameters (Shin et al., 2016). Byra (2021) analyzed these effects with different transfer learning techniques from networks pre-trained on the ImageNet dataset to the new task of cancer detection. They evaluated three standard transfer learning techniques:

- Feature extraction: All layers of the pre-trained network except for a final dense layer replaced for binary classification are frozen and unable to be trained.
- Final layers fine-tuning: All layers of the pre-trained network except for the last CNN block or blocks are frozen and unable to be trained.

- Full fine-tuning: All layers in the pre-trained network are allowed to be trained on the new dataset. In this case, the pre-trained network only acts as a starting point for the parameters of the network.

Using ResNet 101, a popular pre-trained network, on a public dataset of just 647 images, they achieved AUC scores of .903, .934, and .916, respectively, for the above techniques. These results suggest that the final layers are most important to new classification tasks.

Data Augmentation

A straightforward solution to the lack of available training data is to create more images. Instead of duplicating random images like bootstrapping techniques, data augmentation creates new noise by creating copies of randomly selected images with noise introduced. This method has two additional benefits. We train the model to be more robust by invariant to the image's translations by introducing random noise (Taylor et al., 2017). This invariance is particularly useful in medical imaging when lesions are often of different shapes, sizes and located in different regions. Augmentation can also be used to balance the training data with a heavy class imbalance, a common issue in medical imaging (Tanaka et al., 2019).

Although there seems to be little literature on the effects of data augmentation regarding breast ultrasounds. Shorten and Khoshgoftarr (2017) surveyed data augmentation techniques. They found that even simple data augmentation techniques such as flipping, rotation, translation, and cropping helped to lower error rates significantly. However, they warn of the safety of these techniques. For certain domains, these geometric transformations may change the image's label.

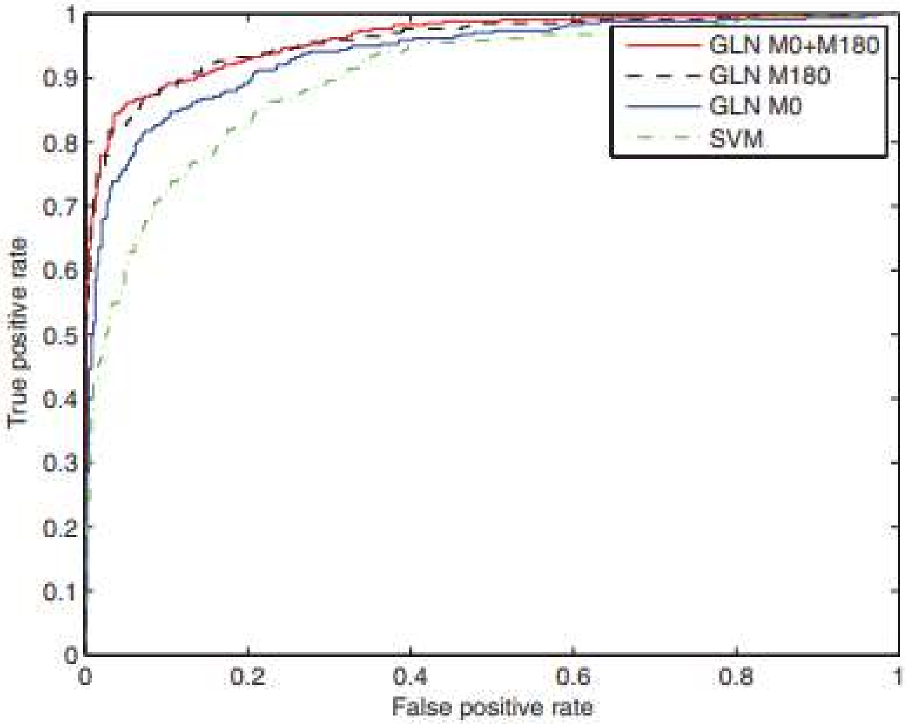
Results

The models reviewed in the literature show that deep learning can be applied to breast cancer detection in ultrasound images with highly accurate results. For example, Saxena (2020) trained MobileNet, commonly used on lightweight devices such as phones and tablets, to achieve precision and recall of .97 and .98, respectively, on the commonly used BUSI public dataset.

Figure 2 shows the receiver operating characteristic (ROC) curves Han et al. (2017) found when applying a mixture of GoogleNet and support vector machines. The ROC is a standard measure used in the literature to gauge classification accuracy. The curve plots a model's true positive rate (TPR) vs. false positive rate (FPR) at various thresholds for the predicted probability.

Figure 2

GoogleNet ROC curve (Han et al., 2017)



Little research has been completed to compare the performance of deep learning algorithms and trained radiologists. Of the literature reviewed, only one compared the results of their deep learning algorithm with the BI-RADS score from a trained radiologist. The research conducted by Kim et al. (2021) found that deep learning drastically improved the AUC from .61 to .89 compared to radiologists.

Limitations and Considerations

Despite the promising results in a research setting, CAD systems have struggled to perform sufficiently in a clinical setting. Some studies even suggest that CAD systems may be increasing the false negative rates of radiologists (Pisano, 2020). Because of the data dependency issues discussed above, deep learning algorithms are highly prone to overfitting (Lee et al., 2017). When applying to new environments, separate patient populations and imaging devices may be to blame for these inconsistencies.

Deep learning models also suffer from a lack of interpretability. The complex features found in the images may be necessary to the algorithms but are hard to decipher for humans (Razzak et al., n.d.). Because of the lack of interpretability, it is often hard to discover where the algorithm performs poorly without a deep understanding of the architecture. Many hospitals and clinics do not have someone with this expertise, so problems may go unresolved.

Conclusion

Deep learning algorithms look to be potentially successful in aiding radiologists to detect breast cancer in ultrasounds. However, it may be hard to develop a trained network that performs well in a clinical setting due to data-dependent algorithms. Fortunately, the data supplied from Mayo Clinic Health System will represent their patient population and be collected using their imaging devices. A highly valuable CAD system is achievable when we couple this with the performance-improving techniques described above.

Chapter 3: Methodologies

Introduction

As described in Chapter 1, one overall goal of this project is to apply deep learning models and advanced training techniques to classify BUS images. This chapter describes the methodology used, including the data acquired and used for training, the applied model architectures, training procedure, and techniques to increase performance.

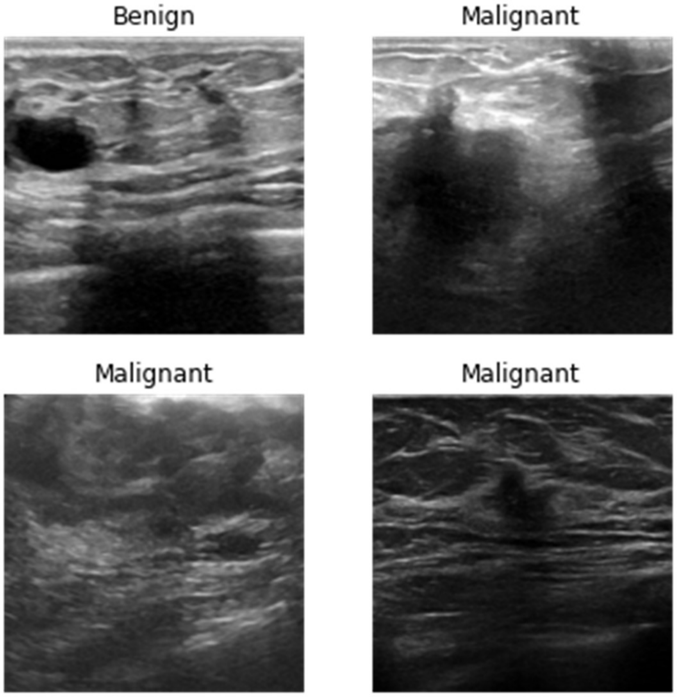
Data Description

At the start of this project, the goal was to use only images supplied by Mayo Clinic Health System. Unfortunately, with the strict protections around highly confidential patient data, we have been unable to successfully gain access to images and their pathologies at the time of writing. Two public datasets were combined to create the dataset for training to provide some results and an overall strategy for classification.

The final dataset consists of 810 images classified as either benign or malignant lesions. The average resolution of the original images is 487×594 pixels. Figure 3 shows a sample of some training images and their classifications.

Figure 3

BUS Image Training Batch



The final dataset was split into three subsets. A set of 567 images was used for training, a set consisting of 121 images was used for validation at the end of each training step, and a hold-out test set containing 122 images was used to see how well the models generalized to unseen data. The class balance is consistent between all three datasets, within ± 6%, with the smallest percentage of benign images being 64% and the highest 70% approximately.

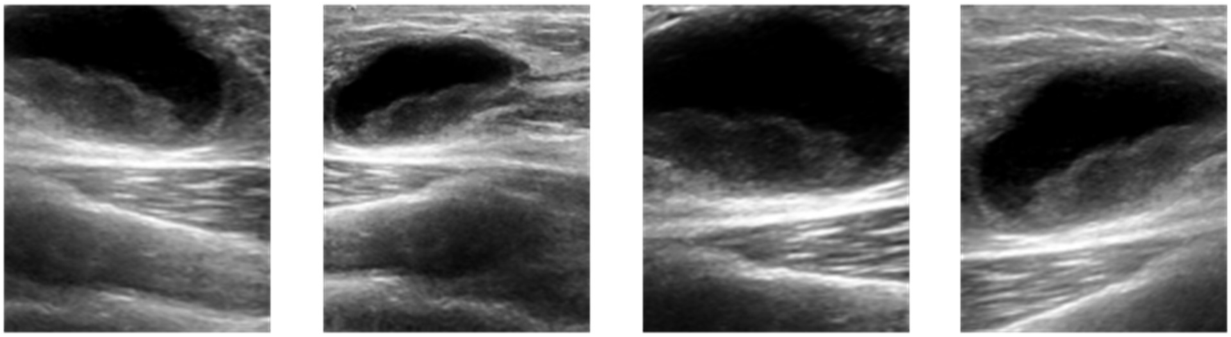
Data Augmentations

A common problem with deep learning architectures is their ability to memorize instead of "learning," making them unable to generalize to new data. With the small amount of training data, it became essential to incorporate techniques to reduce overfitting. By randomly selecting images during each training step to be altered via rotations, reflections, added noise, Etc., we can help the models become more robust. We implemented two different augmentation strategies to help avoid overfitting on our BUS images.

The first augmentation strategy uses geometric transformations, image resizing, and zooming. The geometric transformations used on our dataset could not be chosen without some considerations. Depending on the transformations applied, the final classification may be altered, or the image may no longer make sense in the domain of BUS images. For example, BUS images are taken with the layer of skin towards the top and bone or deep tissue towards the bottom. If we were to flip the image vertically, we would be introducing a transformation that may not be appropriate where the skin was shown beneath bone and dense tissue. Therefore, we have limited transformations only to introduce some image rotation by a slight angle, less than ten degrees in either direction, zooming the image up to 1.5x magnification, and randomly cropping the image to 224×224 pixels. Figure 4 shows the results of these augmentations when applied to an image from our dataset.

Figure 4

Data Augmentation of BUS Images



The second data augmentation method utilized was mixup (Zhang et al., 2017). This method is a data-agnostic approach to augmentation and does not suffer from relying on domain knowledge to apply relevant transformations. According to Zhang et al., this data augmentation is also very powerful when the labels are not entirely accurate. This was extremely important for our BUS images since many of the images have been classified by a human, which we know are prone to errors.

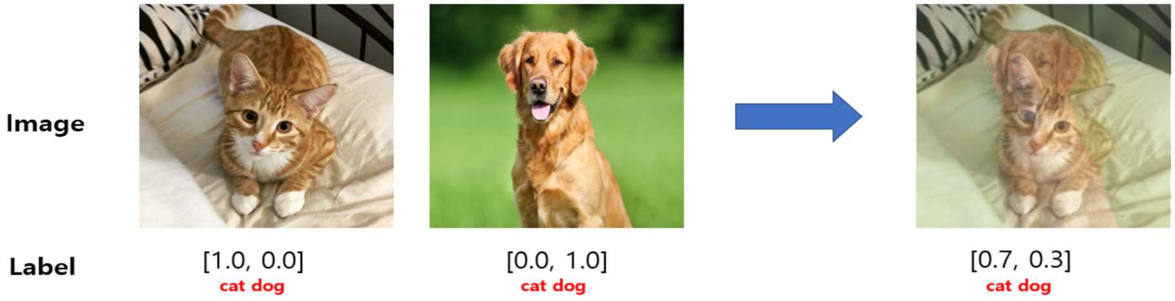
To perform mixup, we generate a new image by randomly selecting two images, x_i, and x_j, and their labels, y_i, and $y_{j,}$ which have been encoded, in our case 0 for benign and 1 for malignant. These images and labels are then combined linearly based on a randomly chosen weight, λ. Figure 5 shows how the new images, $\hat{x}$, and labels, $\hat{y}$, are generated:

Figure 5

Mixup data augmentation

$$\boxed{\begin{aligned}\hat{x} &= \lambda x_i + (1-\lambda)x_j,\\ \hat{y} &= \lambda y_i + (1-\lambda)y_j,\end{aligned}}$$

where $\lambda \in [0,1]$ is a random number



The resulting image may contain essential features of malignant and benign lesions, like the image above with the overlayed cat and dog. The combined image may be consistent with many real-life hard to classify lesions, exhibiting multiple features. The intent was that the model will be more robust by learning which features are more important to a benign or malignant lesion.

Models

Because it is not our intent to create a novel deep learning architecture but apply current state-of-the-art deep learning models, we chose three different well-known architectures to use and analyze. These three architectures Virtual Geometry Group (VGG) neural networks, residual neural networks (ResNets), and densely connected convolutional networks (DenseNets) have become reliable models for nearly all computer vision tasks. Therefore, these models offer a lot of potential in to understanding the capabilities of deep learning for BUS classification. Furthermore, DenseNet, may be especially important for medical imaging because the architecture allows the input to be passed to all layers in the network, enabling small, detailed features to also remain relevant to the final classifications (Iandola, 2014).

Although other deep learning architectures for classification exist, such as visual transformers (ViT), these architectures are still in their infancy, and visualization and understanding of their output are still unexplainable (Khan et al., 2021). Because of this, we chose to focus our analysis on the CNNs mentioned above. Table 2 shows the complete list of models analyzed.

Table 2

Deep learning models for BUS Images

Model	Number of Parameters (millions)	Depth	Approx. Size (MB)
VGG-16	15.25	16	528

Model	Number of Parameters (millions)	Depth	Approx. Size (MB)
VGG-19	20.56	19	549
ResNet-18	11.70	18	44
ResNet-34	21.83	34	83
ResNet-50	25.62	50	98
ResNet-101	44.61	101	171
ResNet-152	60.25	152	232
DenseNet-121	8.00	121	31
DenseNet-169	14.20	169	55
DenseNet-201	20.07	201	77

Training Procedure

Training these models requires lots of computing time and resources, making it hard to analyze how small changes to the model parameters affect performance. To reduce the required training time, we utilized multiple techniques which have been shown to increase the rate at which these models learn.

Transfer Learning

Starting from an initial randomized point for our models can result in inconsistent training times and performance. Instead, we can choose to use starting parameters for our model from pre-trained models. This method known as transfer learning uses the model weights and biases from a model trained on a different task and utilizes them as an initial starting point for our problem. We began our training with the weights and biases for the models in Table 2 trained on the ImageNet dataset, described in Chapter 2.

Each model's final fully connected layer was removed since it was trained to output classifications for a separate task. A new fully connected layer with two outputs, known as the head of the model for our binary classification problem, was added. Because the initial weights for this layer must be randomized, we began training by freezing the entire model, except for this layer. Frozen layers of a model cannot be updated during training. By leaving the head unfrozen this ensured only the weights and biases for the new classification layer were updated during the beginning of training. After updating this layer for the set number of epochs, the entire model was unfrozen, and all layers were updated during the rest of the training.

Discriminative Learning Rates

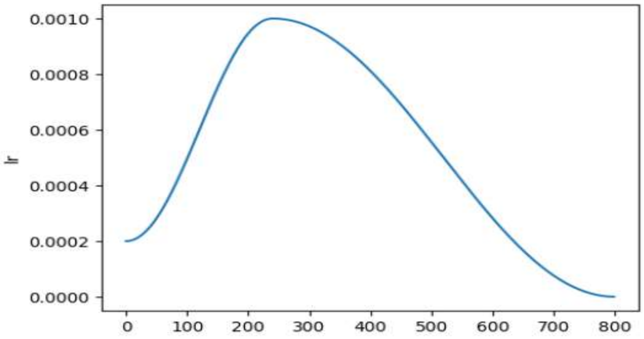
The early layers of our model will be used to detect simple features in our images, such as edges, curves, and corners. These simple features are present in almost all images. Because these layers will begin training with their pre-trained weights, they should require little change. However, as we move deeper through the network layers, the model parameters will require more change to reach an optimal solution because the complex features found for ImageNet will be much different from those for BUS classification. Therefore, it does not seem reasonable that all layers should be trained at the same learning rate. We used discriminative learning, which sets a different learning rate for different depths of our networks. This ensures that our deeper layers and our new head train much faster than the early layers, which should only require minor updates. Ultimately, this should reduce our training time as the model learns the features important to BUS classification more quickly.

Scheduler

It is often the case that the initial learning rate is not optimal throughout all the training. For example, we may find a local minimum in our loss function, and our learning rate may not allow us to leave this local area of our loss function. For our models, we implement cosine annealing, following the 1-cycle policy defined by Smith (2018). Many other schedulers exist, but the 1-cycle policy performs well as a universal scheduler and greatly reduces training time. The 1-cycle policy works by increasing the learning rate from the starting learning rate to a maximum learning rate and then lowering it again to zero over one cycle, typically chosen to be slightly less than the total training epochs as shown in Figure 6.

Figure 6

Learning Rate Scheduler



By maximizing the learning rate during the middle of training, we hope to prevent the model from landing in a steep local minimum.

Hyperparameter Tuning/Optimization

Our models rely on many parameters, and finding the optimal combination of these parameters is likely not achieved by random choice. To find the optimal choices for each parameter, we used the Python library Optuna. Using the Tree-structured Parzen Estimator (TPE) algorithm to search our defined parameter space. TPE was chosen since it has been shown to provide consistent convergence to an optimal solution unlike other methods such as grid or random searches. We select hyperparameters and run multiple trials with each model to reach the optimal results using this method.

Given the imbalance of classes in our dataset, we choose to maximize each model's Area Under Curve score (AUC). AUC is the probability that our models given two images, one benign and one malignant, will classify the malignant image with a higher probability of malignancy. We required at least 15 trials to run to completion, but we stopped any trials early where the AUC on the validation set was below the median AUC after five epochs. The search space used for the TPE optimization is shown in Table 3.

Table 3

Hyperparameter Search Space

Hyperparameter	Min. Value	Max. Value	Distribution
Batch Size	8	64	Uniform
Epochs Frozen	0	25	Uniform
Weight Decay	1e-05	1e-01	Log Uniform
Base Learning Rate	1e-05	1e-01	Log Uniform
Beginning Momentum	.8	.999	Uniform
Min. Momentum	.8	Beginning Momentum	Uniform
Final Momentum	Min. Momentum	.999	Uniform

Chapter 4: Results

Introduction

This chapter reviews the results and findings of our models and training procedure. The models are assessed based on their performance on the validation set of 121 images. Furthermore, we analyze the effects of mixup to avoid overfitting, compare the results of our top-performing model to a trained radiologist, and offer an interpretation into how our models classify images.

Findings

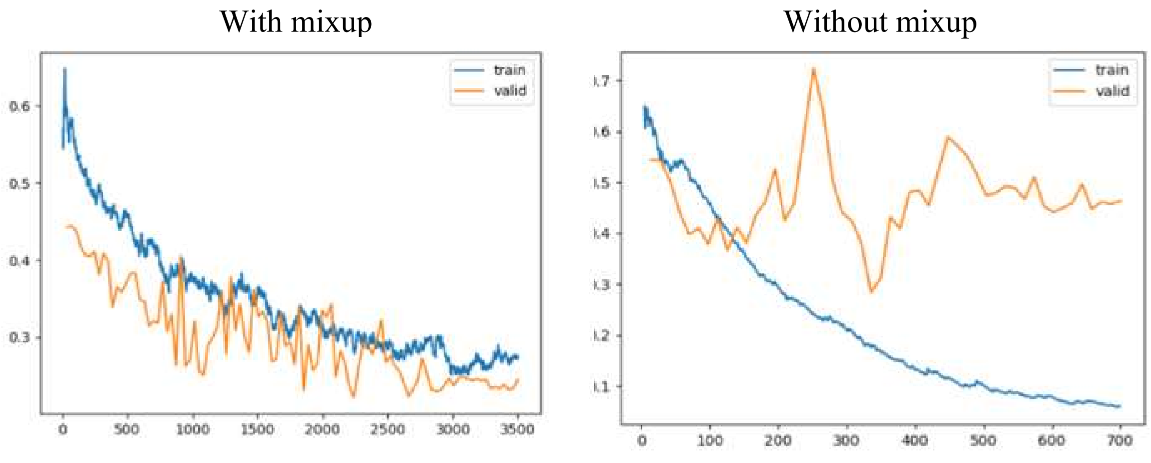
Each model was trained for one-hundred epochs, or training steps, using the training procedure defined in Chapter 3. The model was evaluated on our validation dataset at the end of each epoch. The results from the validation dataset were used to select the final model. The following sections describe the analysis of our training and results.

Effects of Augmentations

During training, the validation loss is usually higher than the training loss because the validation data is not used to update the model weights. However, a strange pattern emerged when training models on the BUS images following our training procedure. The validation loss remained lower than the training loss for nearly all epochs for our models. The training and validation loss for similar ResNet-34 models, only the batch size has changed, with and without mixup applied, are shown in Figure 7.

Figure 7

ResNet-34 Training Comparison using mixup


With mixup

Without mixup

Mixup was highly effective in avoiding overfitting, especially given the small training data. Essentially, by synthesizing more complicated ultrasounds during training, unaltered images in the validation were more effortlessly classified by our models. By increasing the difficulty of our problem during training, we made the most of our small training data and significantly increased our performance.

Model Performance

The classification results for each model displayed in Chapter 3 were obtained and shown below in Table 4. The AUC, precision, recall, F1-score, and accuracy are displayed for each model. Our models defined malignant lesions as positive cases. Therefore, recall and precision here refer to the ability of our model to accurately classify all malignant lesions and the power of a malignant classification, respectively.

Table 4

Deep Learning Results

Model	AUC	Precision	Recall	F1 Score	Accuracy
VGG-16	.963123	.90625	.828571	.865672	.92562
VGG-19	0.95814	0.837838	0.885714	0.861111	0.917355
ResNet-18	0.968771	0.852941	0.828571	0.84058	0.909091
ResNet-34	0.959468	0.794872	0.885714	0.837838	0.900826
ResNet-50	0.968771	0.882353	0.857143	0.869565	0.92562
ResNet-101	0.967774	0.815789	0.885714	0.849315	0.909091
ResNet-152	0.971096	0.846154	0.942857	0.891892	0.933884
DenseNet-121	0.968439	0.861111	0.885714	0.873239	0.92562

Model	AUC	Precision	Recall	F1 Score	Accuracy
DenseNet-169	0.966113	0.848485	0.8	0.823529	0.900826
DenseNet-201	0.965116	0.888889	0.914286	0.901408	0.942149

We can see that all the state-of-the-art models achieved similar results. These results suggest we are likely reaching some limitations with our dataset as we would expect some differences between the shallower and deeper models.

Model Choice

While many models have comparable AUC values, DenseNet-201 seems to best balance the difference between precision and recall giving us the maximum F1 score. We will use this model to evaluate our performance in the following sections. However, this model still struggled to classify some of the images in our validation set successfully. Table 5 shows the confusion matrix for our validation set. We can see that the model made ten separate incorrect predictions on the validation data set. This would have resulted in five patients missing the critical care they would need and five receiving care and likely invasive procedures unnecessarily.

Table 5

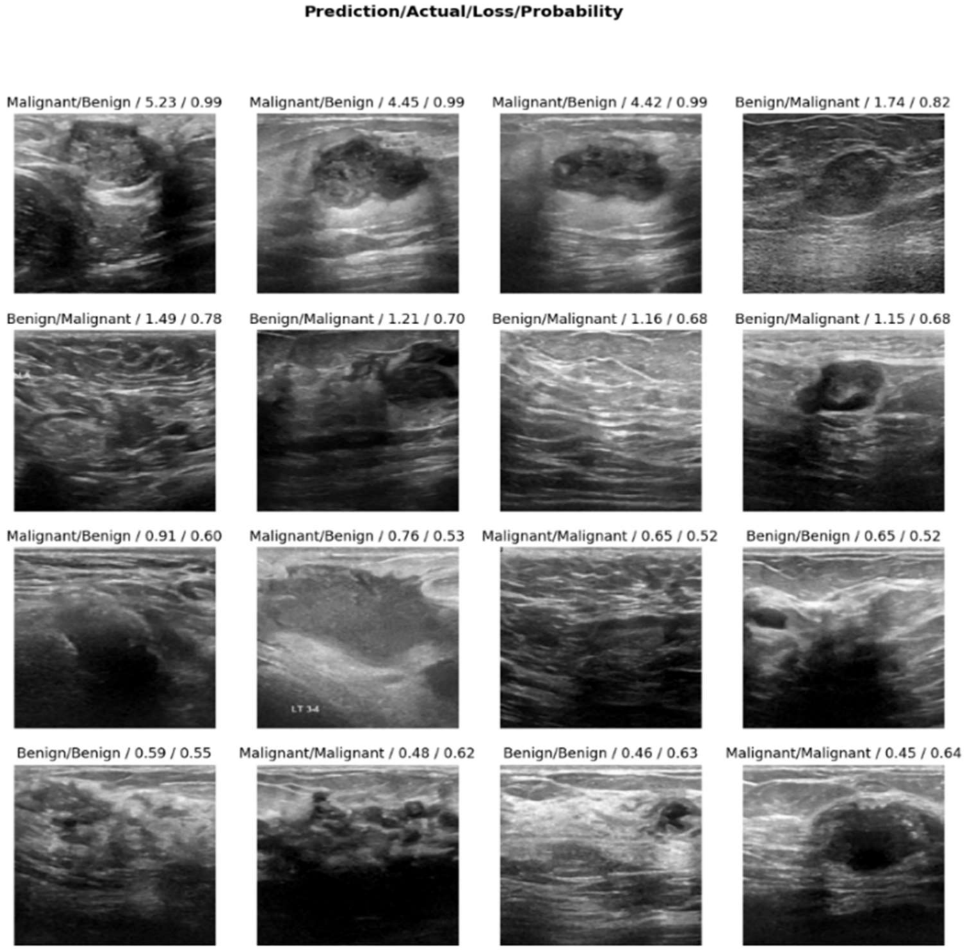
DenseNet-201 Validation Set Confusion Matrix

	Predicted Benign	Predicted Malignant
Actual Benign	81	5
Actual Malignant	5	30

Figure 8 shows these ten misclassifications and six other images that most contributed to the validation loss. We can see that aside from the misclassifications seen in the confusion matrix, six images were predicted correctly, but the confidences of those predictions by the model were low.

Figure 8

DenseNet-201 Top 16 Losses


Prediction/Actual/Loss/Probability
Malignant/Benign / 5.23 / 0.99
Malignant/Benign / 4.45 / 0.99
Malignant/Benign / 4.42 / 0.99
Benign/Malignant / 1.74 / 0.82
Benign/Malignant / 1.49 / 0.78
Benign/Malignant / 1.21 / 0.70
Benign/Malignant / 1.16 / 0.68
Benign/Malignant / 1.15 / 0.68
Malignant/Benign / 0.91 / 0.60
Malignant/Benign / 0.76 / 0.53
Malignant/Malignant / 0.65 / 0.52
Benign/Benign / 0.65 / 0.52
Benign/Benign / 0.59 / 0.55
Malignant/Malignant / 0.48 / 0.62
Benign/Benign / 0.46 / 0.63
Malignant/Malignant / 0.45 / 0.64

It is also possible for deep learning architectures to eventually learn the validation data indirectly, making the results biased. To better understand our model's true performance, we can look at the predictions on data that the model has never seen. The confusion matrix for the test set is shown below in Table 6.

Table 6

DenseNet-201 Test Set Confusion Matrix

	Predicted Benign	Predicted Malignant
Actual Benign	76	2
Actual Malignant	7	37

Our model did well even on new data, with comparable results to our validation results. The model misclassified seven malignant images, but our test set contained more malignant lesions than our validation set. Overall, the model resulted in a precision of .9487 and a recall of .8409 on unseen data.

Comparison to Trained Professional

Since this project aims to help trained radiologists make diagnoses, it is essential to understand if these models would have performed similar or better to a trained radiologist. Of our test set of 122 images, 26 of these were also annotated by Dr. Richard Ellis of Mayo Clinic – La Crosse, a trained radiologist specializing in breast imaging. He assessed the BI-RADS score and gave an expected pathology to classify these images. Dr. Ellis correctly classified 23 of the 26 images shown in Table 7. Of the images misclassified, two were given a BI-RADS score of 3 and 4. These BI-RADS scores are the most misclassified and problematic for radiologists as they are the intermediate scores for whether a biopsy is performed.

Table 7

Confusion Matrix for Trained Radiologist

	Predicted Benign	Predicted Malignant
Actual Benign	17	0
Actual Malignant	3	6

When evaluating our model on this small test set, our model slightly outperformed Dr. Ellis. The model only incorrectly predicted the pathology of two of the images. These two images were also two of the images Dr. Ellis had misclassified. Table 8 shows the confusion matrix of our model on this dataset.

Table 8

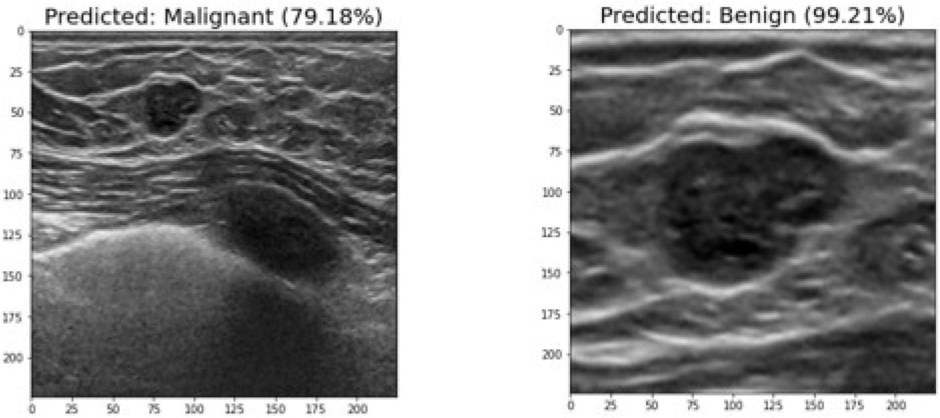
DenseNet-201 Confusion Matrix for Comparison

	Predicted Benign	Predicted Malignant
Actual Benign	17	0
Actual Malignant	2	7

When investigating the image that our model successfully classified but Dr. Ellis did not, something interesting was found. Our model was trained to take the entire BUS image, but Dr. Ellis provided annotations for each image outlining the region of interest around the lesion. When we crop the image to the area identified by Dr. Ellis, we obtain a benign classification from our model. Figure 9 shows how our prediction changes for the cropped image.

Figure 9

Cropped image prediction comparison



Therefore, our model appears to be identifying areas of the image outside of the typical region of interest in making its predictions. Understanding how these predictions are made may provide new and valuable information to radiologists in their understanding of malignant cells.

Interpretation

When making a classification dealing with patient outcomes, radiologists must trust the models' classification. Deep learning architectures often operate as black boxes, making it hard to interpret why the classifications were made. To help interpret these models, we implemented local interpretable model-agnostic explanations (LIME). LIME creates synthesized images from a given image, using perturbations in the localized area of the image to see how that area affects the probability associated with a classification. The areas of most importance are then found using linear models such as weighted linear regression (Ribeiro et al., 2016). LIME was applied to three different malignant images from our dataset. To understand what our model was learning, Dr. Ellis evaluated each image and the LIME results. Each image and LIME results are displayed below in Figures 10 through 12.

Figure 10

Malignant lesion with Cystic Structure

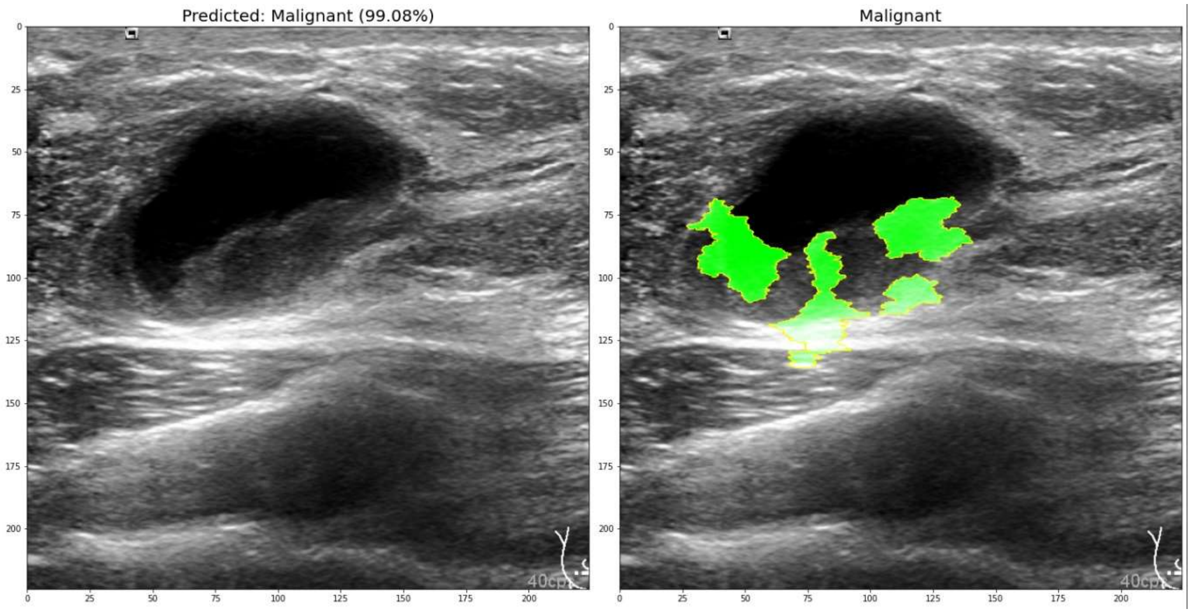


Figure 11

Malignant lesion with non-uniform internal mass echogenicity

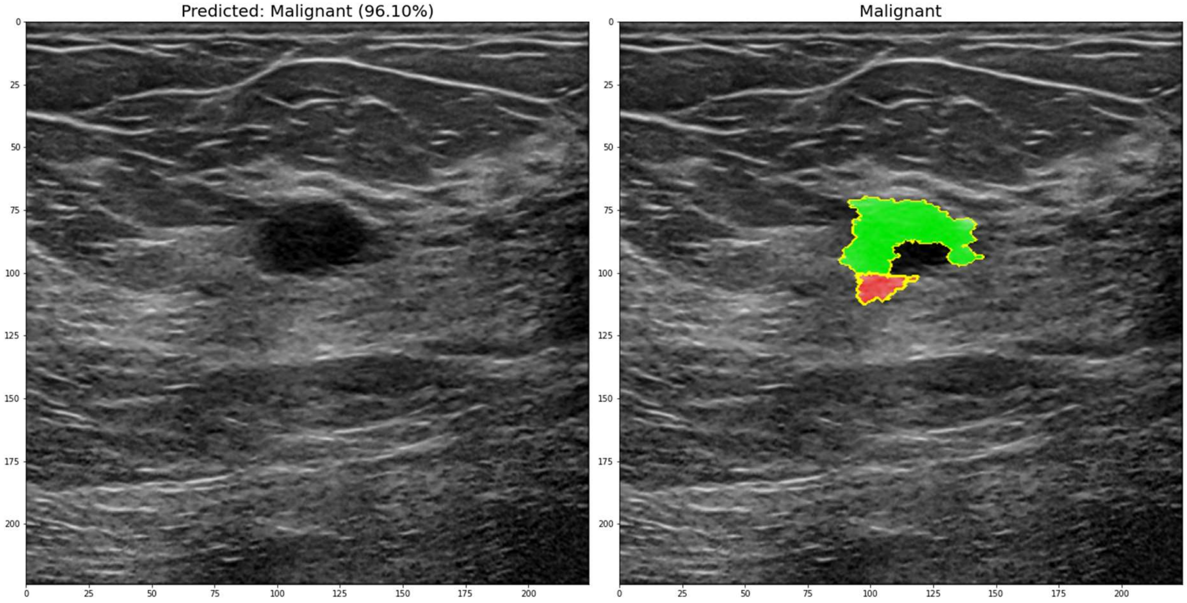
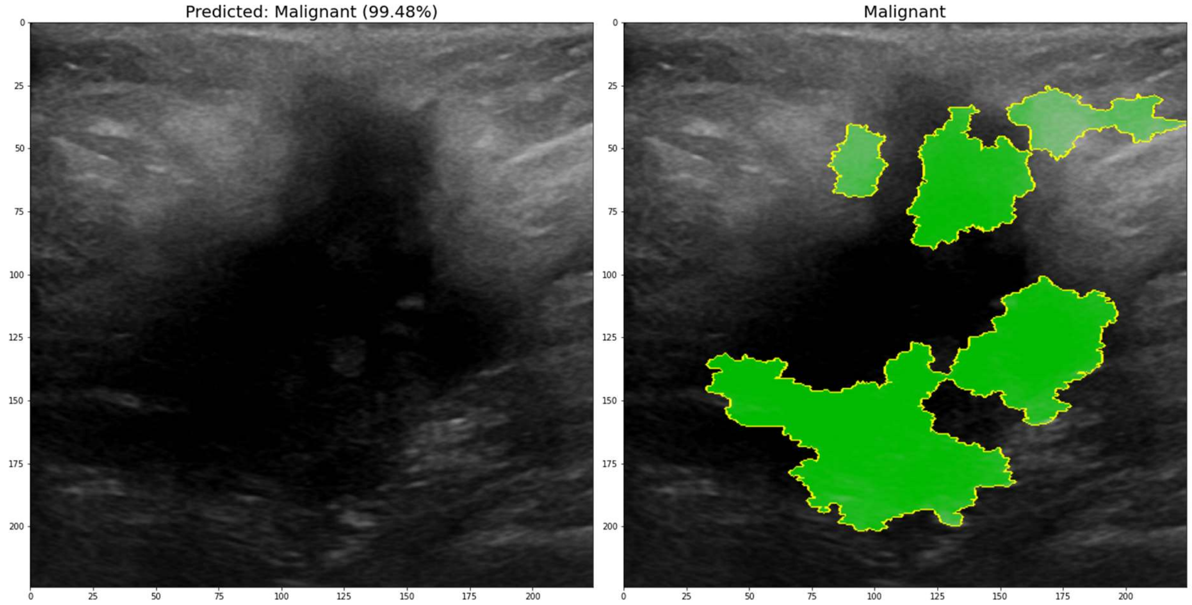


Figure 12

Malignant lesion with features in multiple zones



The original image is displayed on the left of each figure, with our model's prediction. For each figure, the image on the right shows the areas that most increase or decrease the probability of the malignant classification in green and red, respectively. Only the top five areas with statistical significance are shown for each image. Less than five areas are highlighted for some images, such as Figure 10 and Figure 11, if five significant areas could not be identified.

According to Dr. Ellis, the model does appear to be highlighting malignant features. He noted that the model appeared to identify cystic structure in Figure 10, non-uniform internal mass echogenicity in Figure 11, and regions of likely malignant features in the peripheral, marginal, and internal zones of Figure 12.

Dr. Ellis did note, however, that there were other malignant features in each image that the model had not recognized. There are two potential reasons LIME did not identify these areas.

The first is that the DenseNet-201 model had not learned these features during training. The other could be insufficient statistical evidence of their importance, or they were not in the top five regions of importance. Regardless, this approach gives insight to our model's predictions and helps give credibility to the potential for deep learning architecture in a clinical setting.

Conclusion

The findings show that state-of-the-art deep learning architectures can be used to classify BUS images successfully. Even with a small amount of training data, utilizing strong data augmentations such as mixup, we can help the models to generalize to new unseen data. The ability to generalize helped our DenseNet-201 model outperform a trained radiologist. Although the test set used for comparison was small and contained few BI-RADS 3 and 4 images, the model and training procedure show promising results towards providing Mayo Clinic Health System with a successful CAD system.

Chapter 5: Discussion

Introduction

This study aimed to apply deep learning models to aid in the classification of BUS to create a more standardized approach to classifying BUS images. This chapter will include a summary of our research findings related to the effectiveness of deep learning architectures and training techniques that may help these models in a clinical setting. Also included in this chapter is a discussion of the findings, suggestions for future research, and a brief closing summary.

Summary of Findings

Given the small number of training images, all models saw approximately the same performance as measured by AUC. Still, we confirmed that our models were not solely memorizing the data. We avoided overfitting our small sample using mixup and better helped our model become more robust showing success on unseen data. DenseNet-201 showed a better balance between precision and recall, resulting in the maximum F1-score obtained. When evaluating this model on our test set, this model performed well, with comparable performance to a trained radiologist on a small subset of annotated data. Using LIME, we were able to show that an essential part of these classifications was the fact that the model was able to learn complicated properties of the lesion also identified by a trained radiologist in a short time. These findings suggest that current deep learning architectures offer strong potential for aiding radiologists in a clinical setting.

Discussion

At the onset of the project, our goal was to primarily use data supplied by Mayo Clinic Health System. However, our models relied heavily on public data sets due to the sensitivity around patient data and the time it takes to label and annotate the images correctly. Although these images contain the same modality of ultrasounds, the image quality, resolution, and patient population are likely to be different from images supplied solely by Mayo Clinic Health System.

However, the training procedure and data augmentations we have applied are not limited to the current data set. Given enough data from Mayo Clinic Health System, these models could easily be applied to new data and evaluated for their performance. We have seen that the models generalize well to unseen data from the same population. Therefore, it is very promising that we can deliver Mayo an effective CAD system using these models when we have more data from them.

It should also be noted that although our models slightly outperformed a trained radiologist, that is not the primary intent of this project. While we hope to achieve an accurate classifier to improve patient outcomes, our primary aim is to provide a standardized way of aiding providers in assigning a BI-RADS score and overall classification. A crucial part of achieving this goal is to design models that radiologists can trust. Our findings provide a way to give insight into our predictions and further aid radiologists in making standardized decisions by bringing attention to critical areas of the ultrasound image.

Suggestions for Future Research

While significant progress was made to show these models can aid in BUS classifications, the applied deep learning methods still leave room for improvement. Two options are discussed below that fell outside of the scope of this paper.

Although not discussed in this paper, an essential part of the CAD system is the detection and segmentation of lesions in BUS images. The focus has been to create two separate models, one for classification and object detection. The classification models discussed above have been trained on a complete BUS image, which may contain areas of the image not relevant to the classification. For example, beyond the muscle is often too deep for the ultrasound to effectively penetrate. These regions and objects like bone may often appear as dark lesions exhibiting malignant features. As result our models may have an increased number of false positives because of these regions. Developing a pipeline of models may be more beneficial. If we first use a model for object detection to generate a proposed region for the lesion, we can crop the image to focus more directly on the lesion and surrounding area. The classification model can then be trained on these cropped images. These cropped images may help the model learn more quickly what features in the image are essential to the classification and avoid false areas of interest.

The second option for future research aims to solve another problem experienced so far. One of the significant challenges with this project has been gaining access to the data, and the time it takes to annotate and label each image. The second, more challenging solution may be using semi-supervised learning algorithms. Semi-supervised algorithms can utilize unlabeled data in their training and develop robust learning from a small number of labeled images. The bottleneck associated with labeling medical images could be greatly reduced by incorporating these algorithms for classification or data labeling. This would allow for large batches of images to quickly be applied to supervised learning algorithms, improving the results of these algorithms and their use in clinics.

Conclusion

Current state-of-the-art deep learning architectures show considerable potential to aid radiologists in a clinical setting. However, the sensitivity of the data in the medical field continues to make it challenging to achieve the full potential of these models. Even with this challenge, we have shown that using techniques during training and powerful data augmentations, we can build effective models with value to radiologists. As we continue to get access to more data and progress to other phases of this project, the use of deep learning looks very promising. Eventually, we will be able to develop a completely functional CAD system fully utilizing these techniques, bringing value to radiologists across the Mayo Clinic Health System.

References

Breast cancer now most common form of cancer: WHO taking action. (2021, February 3). Retrieved December 5, 2021, from https://www.who.int/news/item/03-02-2021-breast-cancer-now-most-common-form-of-cancer-who-taking-action

Breast Imaging Reporting & Data System. (n.d.). Retrieved December 5, 2021, from https://www.acr.org/Clinical-Resources/Reporting-and-Data-Systems/Bi-Rads

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

Chiao, J.-Y., Chen, K.-Y., Liao, K. Y.-K., Hsieh, P.-H., Zhang, G., & Huang, T.-C. (n.d.). *Detection and classification the breast tumors using mask R-... : Medicine*. Retrieved October 28, 2021, from https://journals.lww.com/md-journal/Fulltext/2019/05100/Detection_and_classification_the_breast_tumors.4.aspx

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (n.d.). *ImageNet: A Large-Scale Hierarchical Image Database*. 8.

Gómez Flores, W., Pereira, W. C. de A., & Infantosi, A. F. C. (2015). Improving classification performance of breast lesions on ultrasonography. *Pattern Recognition*, *48*(4), 1125–1136. https://doi.org/10.1016/j.patcog.2014.06.006

Huynh, B., Drukker, K., & Giger, M. (2016). MO-DE-207B-06: Computer-Aided Diagnosis of Breast Ultrasound Images Using Transfer Learning From Deep Convolutional Neural Networks. *Medical Physics*, *43*(6Part30), 3705–3705. https://doi.org/10.1118/1.4957255

Iandola, F., Moskewicz, M., Karayev, S., Girshick, R., Darrell, T., & Keutzer, K. (2014). DenseNet: Implementing Efficient ConvNet Descriptor Pyramids. *ArXiv:1404.1869 [Cs]*. http://arxiv.org/abs/1404.1869

Khan, S., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S., & Shah, M. (2021). Transformers in Vision: A Survey. *ArXiv:2101.01169 [Cs]*. http://arxiv.org/abs/2101.01169

Kim, S.-Y., Choi, Y., Kim, E.-K., Han, B.-K., Yoon, J. H., Choi, J. S., & Chang, J. M. (2021). Deep learning-based computer-aided diagnosis in screening breast ultrasound to reduce false-positive diagnoses. *Scientific Reports*, *11*(1), 395. https://doi.org/10.1038/s41598-020-79880-0

Lee, J.-G., Jun, S., Cho, Y.-W., Lee, H., Kim, G. B., Seo, J. B., & Kim, N. (2017). Deep Learning in Medical Imaging: General Overview. *Korean Journal of Radiology*, *18*(4), 570–584. https://doi.org/10.3348/kjr.2017.18.4.570

Mendelson, E., Böhm-Vélez, M., & Berg, W. (2013). *ACR BI-RADS Ultrasounds*. American College of Radiology.

Pisano, E. D. (2020). AI shows promise for breast cancer screening. *Nature*, *577*(7788), 35–36. https://doi.org/10.1038/d41586-019-03822-8

Razzak, M. I., Naz, S., & Zaib, A. (n.d.). *Deep Learning for Medical Image Processing: Overview, Challenges and Future*. 30.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). “Why Should I Trust You?”: Explaining the Predictions of Any Classifier. *ArXiv:1602.04938 [Cs, Stat]*. http://arxiv.org/abs/1602.04938

Saxena, A. (2021). Comparison of two Deep Learning Methods for Classification of Dataset of Breast Ultrasound Images. *IOP Conference Series: Materials Science and Engineering*, *1116*(1), 012190. https://doi.org/10.1088/1757-899X/1116/1/012190

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data*, *6*(1), 60. https://doi.org/10.1186/s40537-019-0197-0

Smith, L. N. (2018). A disciplined approach to neural network hyper-parameters: Part 1 -- learning rate, batch size, momentum, and weight decay. *ArXiv:1803.09820 [Cs, Stat]*. http://arxiv.org/abs/1803.09820

Sun, Y.-S., Zhao, Z., Yang, Z.-N., Xu, F., Lu, H.-J., Zhu, Z.-Y., Shi, W., Jiang, J., Yao, P.-P., & Zhu, H.-P. (2017). Risk Factors and Preventions of Breast Cancer. *International Journal of Biological Sciences*, *13*(11), 1387–1397. https://doi.org/10.7150/ijbs.21635

Tanaka, H., Chiu, S.-W., Watanabe, T., Kaoku, S., & Yamaguchi, T. (2019). *Computer-aided diagnosis system for breast ultrasound images using deep learning*. *64*(23), 235013. https://doi.org/10.1088/1361-6560/ab5093

Taylor, L., & Nitschke, G. (2017). Improving Deep Learning using Generic Data Augmentation. *ArXiv:1708.06020 [Cs, Stat]*. http://arxiv.org/abs/1708.06020

U.S. Breast Cancer Statistics. (2021, February 4). Breastcancer.Org. https://www.breastcancer.org/symptoms/understand_bc/statistics

Vijayarajeswari, R., Parthasarathy, P., Vivekanandan, S., & Basha, A. A. (2019). Classification of mammogram for early detection of breast cancer using SVM classifier and Hough transform. *Measurement*, *146*, 800–805. https://doi.org/10.1016/j.measurement.2019.05.083

Zhang, H., Cisse, M., Dauphin, Y. N., & Lopez-Paz, D. (2018). mixup: Beyond Empirical Risk Minimization. *ArXiv:1710.09412 [Cs, Stat]*. http://arxiv.org/abs/1710.09412

Zhang, Z., Li, Y., Wu, W., Chen, H., Cheng, L., & Wang, S. (2021). Tumor detection using deep learning method in automated breast ultrasound. *Biomedical Signal Processing and Control*, *68*, 102677. https://doi.org/10.1016/j.bspc.2021.102677

Appendix A: Code

The code for this project was created using a mixture of Python and Jupyter notebooks. The Python scripts and notebooks can be found here: https://github.com/hall4jm/CapstoneBUS.