Localization Effects in Deep Learning Classification of Breast Cancer Ultrasounds

Rachel Kim

University of Wisconsin - La Crosse

DS 785: Capstone

Instructor: Alexander Korogodsky

April 28, 2024

Abstract

Early detection of malignancy is a critical factor to effectively treating breast cancer. As radiologist determination of breast cancer ultrasound (BUS) images for malignancy is highly variable across providers, there is a need to develop computer-aided diagnosis (CAD) tools that can improve consistency and reliability of BUS diagnosis. Existing literature on the subject has shown promising results in the direction of creating a deep learning model to predict the malignancy of a BUS lesion. This study is a part of a larger ongoing research initiative between the University of Wisconsin-La Crosse and Mayo Clinic, where creating a CAD tool for BUS image classification to assist in radiologist diagnoses is the primary goal. For this study, various input images are proposed and explored to determine if adding localization information to the deep learning model inputs increases the model's predictive power. The research hypothesis is that adding localization information will improve a model's performance by highlighting the key diagnostic components to the deep learning model. Two pre-trained PyTorch models, ResNet18 and EfficientNetB3, were constructed and trained on three types of custom image inputs including localization information as well as a control of the original BUS image. A combination of private and public BUS images were used, totaling 3,784 images and associated binary masks. The results after training and testing the models on each input were that localization information in a custom input image does increase the predictive performance of a model, with further experimentation needed to identify which specific input is the optimal choice. These results suggest that incorporating localization into a CAD tool for BUS diagnosis could improve the process for both patients and hospitals.

Keywords: Breast cancer, ultrasound, deep learning, localization

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Background

According to the National Breast Cancer Foundation (2024), approximately one out of every eight women are diagnosed with breast cancer within their lifetime. If breast cancer is detected and treated in its early stages, the likelihood of survival is high (Ginsburg et al., 2021). Breast ultrasounds (BUS) are a non-invasive diagnostic tool that gives radiologists a view into a patient's breast tissue for investigating suspicious masses flagged from a mammography, allowing for the early detection of cancer.

Once one or more BUS images are acquired, a radiologist may then diagnose a patient's lesion using a numbered category using a scoring system called Breast Imaging Reporting and Data System, or BI-RADS (American Cancer Society, n.d.). Using BI-RADS, a patient will receive a number from 0-6 indicating the lesion's diagnosis, where a higher number would indicate a larger likelihood of malignancy (see Appendix A).

Although assessing BUS images can be an effective method of diagnosing breast cancer, there is currently wide variability present in how radiologists may read an ultrasound and prescribe further treatment. The diagnosis from a BUS image highly depends on the skills and experience of the provider, methods of operation, and equipment parameters (Dai et al., 2021). If the same BUS image with a BI-RADS score from 3-5 is shown to ten different radiologists, it is likely they may provide different diagnoses, and consequently, different treatment plans.

To assist in the consistency and accuracy of diagnoses, computer-aided diagnosis (CAD) tools have been developed and implemented. Common types of CAD tools for BUS classification today utilize deep neural network methodologies to ingest BUS images with expert-annotated masks and make a prediction as to whether a lesion is benign or malignant

(Bahl, 2022). This study was performed in collaboration with an ongoing research initiative between the University of Wisconsin-La Crosse and Mayo Clinic Enterprise, referred to as the BUS Project. The ultimate goal of the BUS Project is to create a CAD tool to increase the consistency of radiologist diagnoses in the hospital setting. In this study, the addition of localization effects on input images for deep learning models were explored to improve the performance of BUS classification models with a CAD tool in mind.

Statement of the Problem

The diagnosis of BUS images by radiologists can be highly subjective, consequently influencing both a patient's journey and hospital resources. For patients, the process of undergoing diagnosis for breast cancer can be an extremely emotional and anxiety-inducing time. The feelings of loss and distress can alter not only the emotional well-being of patients but also their physical health as well. Communication from the providers regarding diagnostic results dramatically impacts the patient's journey and ability to understand and cope with their state of health, so it is critical that radiologists and oncologists provide the most accurate and complete diagnoses possible (Ciria-Suarez et al., 2021).

In addition to significantly impacting a patient's journey, the financial cost of misdiagnosis is an incentive to hospitals for improving the accuracy of diagnostic readings of BUS images. An estimate from a study by Ong and Mandl of the costs associated with false positive screenings for breast cancer in the United States was nearly $4 billion annually (2015). False positives are fairly common, with roughly half of women participating in annual breast screenings over a 10-year period receiving at least one false positive diagnosis (American Cancer Society, n.d.). Improving the consistency and reducing false positives in BUS readings when

determining whether a breast lesion may be malignant or benign is a worthwhile problem to tackle both for patients and healthcare organizations.

Purpose of the Study

With the motivation rationale to improve the performance of breast cancer diagnosis using BUS images outlined, this study sought to accomplish this task by exploring the effects that incorporating localization information in the input images of a deep learning model would have on the model's predictive power. Several classifier models using PyTorch architecture were constructed and trained on various input images to identify an optimal input. The long-term goal of the BUS Project is to create a CAD tool to implement within hospitals that will assist radiologists in malignancy determination for lesion ultrasounds. Identifying an optimal image input for such a model that will provide robust, consistent results is a critical step in this process that this study aims to accomplish.

Research Questions

When considering an input for a deep learning model that would produce the most consistent and reliable predictions for malignancy in BUS images, several questions were posed at the beginning of the study to guide exploration and research efforts. Before constructing models and testing inputs, it was known that the models to be used would be based on pre-trained PyTorch deep learning models. The optimal image input compared against the standard BUS image as a control was investigated systematically. The following research questions were posed before developing methods:

1. How does varying localization information added to an input image affect the model's predictive performance?

2. Does varying localization information of an input image affect different model architectures in different ways?

Significance of the Study

A reliable deep learning classifier model in breast cancer diagnosis applied as a CAD tool for radiologists has a high potential to make an impact on both patient lives and hospital funding. Creating a deep learning model that ingests BUS images and predicts malignancy with high accuracy is a challenging task, even for modern advanced architectural frameworks. Convolutional neural networks (CNNs), a type of deep learning model, have been able to classify breast cancer ultrasound images with up to 87.5% accuracy (Cao et al., 2019). Constructing a model that can match or exceed these standards using a custom input image with localization and implementing it into a CAD tool could increase standardization and consistency within the breast cancer diagnostic process. Patients could receive fewer false positive diagnoses, consequently reducing unnecessary emotional turmoil and further health degradation. Hospitals would also save potentially billions of dollars by reducing the expenditures associated with false positives and the missed true negatives.

Limitations and Delimitations of the Study

Although this study utilized high-quality, expert-labeled ultrasound and mask images in creating a classifier model, data quantity was one limitation noted. Deep learning models are data-hungry, requiring large amounts of training data to make accurate classifications. A relatively small quantity of grayscale and annotated ultrasound images were provided by the Mayo Clinic, totaling about 275 anonymized ultrasound and mask images each. Additional datasets of public BUS images with associated masks were added to expand the dataset size, totaling 3,784 images, which is still relatively small for training a deep learning model. However,

data augmentation alleviated the overfitting of the deep learning models that resulted from the small size of the dataset. Various transformations applied to the existing training images mimicked natural variability found in medical imaging and artificially expanded the sample size (Lo et al., 2018).

Whilst small datasets sometimes result in poor model performance when applied to new data, there is a cutoff in which adding additional data does not necessarily bring performance improvements. Rajput et al. (2023) suggest that a small dataset for a machine learning model is sufficient when appropriate effect sizes and the overall model accuracy on new data is greater than or equal to 80%. For this study, the data was augmented to add additional training information and overcome its original size limitation, with the final result being a high-performance model that generalizes to new inputs adequately.

Conclusion

In this study, the challenges of constructing a robust deep learning classifier model tailored for BUS images and identifying the effects of localization on the model performance were pursued. Through systematic experimentation, the localization effects and suggested optimal input image type were identified. The resulting findings may be integrated into a future CAD tool to enhance diagnostic consistency within breast cancer screening. Increasing the performance of malignancy prediction in breast cancer ultrasound readings improves the patient experience, as well as reduces the financial burden of the hospital by decreasing the number of misdiagnoses. In subsequent chapters, the review of the literature around deep learning models for BUS classification, data collection and methodology, study results, and next steps of the larger BUS Project are discussed in detail.

Chapter 2: Literature Review

Introduction

In recent years, advances in machine learning have provided new opportunities across fields. Deep learning, a subset of machine learning using artificial neural networks, overtook other types of machine learning methods sometime in the 2010s and remains the most promising framework for identifying complex patterns buried within data (Hao, 2019). Convolutional neural networks (CNNs), specifically, have shown promise as a valuable tool within the medical diagnostic industry. Research on utilizing CNNs as a framework for computer-aided diagnosis (CAD) tools for radiologists has been conducted within the past decade, with some experimentation performed on varying input images to computer vision classifiers.

Because the goal of this study was to identify the effects that incorporating localization information into input images has on a BUS classifier model, exploring the existing research was necessary to first set a foundational understanding and baseline for new experimentation. A review of the literature was performed to assess the past and present landscape of BUS image classification using CNNs to guide this BUS Project study, as well as identifying the gaps in the literature that this study aims to address. Using the online University of Wisconsin-La Crosse Murphy Library and Google Scholar as search engines, the most relevant research and trade articles were pulled for review related to deep learning for BUS imaging, CAD tools, and fusion or enhanced BUS images for breast cancer diagnosis. The findings of the review of literature are now presented in the remainder of this chapter.

Deep Learning for BUS Classification

A CNN performs image classification through a series of convolution and pooling layers, followed by fully connected layers and a predicted output class. There are many articles

summarizing this broad concept from the past decade, with a comprehensive review of these writings by Rawat & Wang (2017). Research regarding the use of CNNs for medical diagnostic purposes has focused on addressing the challenges of determining the best combination of architectural frameworks and inputs into the classifier model while preserving explainability in the model's predictions. Bahl stated that because mammography can often not provide enough information for breast cancer diagnostic purposes, ultrasound screenings have gained popularity (2022). As of this study conducted in 2022, there were only two CAD applications for BUS classification approved by the United States Food and Drug Administration (FDA), but even those two reported low specificity and higher false positives.

For training a CNN from scratch on classifying BUS images, a large amount of annotated images are required to begin producing substantial results. Because this quantity of anonymized, expertly annotated images can be difficult to obtain, transfer learning has become a common practice in the research community. Transfer learning involves using pre-trained image classification networks as a starting model for a different application to save time and resources. Ragb et al. identified several promising pre-trained CNNs for BUS classification, including VGG16, ResNet50, ResNet18, and GoogleNet, and combined these through ensemble learning (2021). This approach using transfer learning and ensembling calculates a prediction for each network and then applies a majority vote classifier to make a final, more robust prediction.

Explainability of CNNs for BUS Classification

Although several deep learning frameworks have proven to have fairly high performance in predicting malignancy from BUS image inputs, explainability has not always been considered. Even if a CNN can predict accurately, the reasoning behind why an algorithm identifies a lesion as malignant or benign must be apparent to radiologists who wish to use this information for

prescribing treatment regimens. A study by Zhang, B. et al. introduced the idea of using convolutional layers to extract the individual BI-RADS features useful for classifying a lesion first, and then passing those features to a fully-connected layer for a final prediction (2021). After a final prediction is made, a clinician could then look back at the features extracted to understand the rationale of how and why the prediction was made. BI-RADS features have the potential to increase a model's predictive power while offering increased explainability, although they have not been juxtaposed with other localization techniques as done in this paper.

Fusion Image Input for Deep Learning Classifiers

With many researchers striving to create a robust classifier CNN model for BUS diagnosis, various input images have already been tested to see their effects on accuracy. Because most of these models are built with transfer learning, a standard input of a 3-channel image is expected by the pre-trained CNNs. Therefore, images that only have a single channel of useful information present an opportunity for researchers to fill in the remaining two channels with information they believe will improve their model's ability to predict malignancy. For instance, the study by Zhang, B. et al. took all images that only contained the grayscale ultrasound information in the first channel, added a second channel of histogram equalization applied to the grayscale content, and a third channel of smoothing applied to the grayscale content (2021). This study reported that there was a performance improvement by adding the simple additional channels, but did not quantify the improvement the fusion input image provided. Similarly, Yap et al. utilized the second and third previously unused channels of the input image of their CNN to add a sharpened grayscale image and contrast-enhanced grayscale image, respectively (2020). The results from the study were stated as inconclusive in regards to the effect that the pre-processed 3-channel image had on the model's performance.

More complex images derived from the original BUS and mask images have also been explored. For example, BI-RADS feature maps were used by Zhang, E. et al. to provide tailored classification information to the CNN (2020). These BI-RADS feature maps were obtained through a distance transformation paired with a Gaussian filter. These BI-RADS feature images from the study include features such as the lesion boundary, angular characteristics, and the average intensities of grayscale surrounding the lesion. This study showed a relatively high accuracy of 94% achieved using this methodology but failed to include multiple lesions from separate angles captured within the datasets used.

A slightly simpler approach than including all BI-RADS features was taken by Daoud et al., where morphological and texture information from the input images were used to classify lesions using a support vector machine (2016). These morphological characteristics, such as perimeter, roundness, and tumor area, allowed the classifier to achieve a high accuracy of 98.2% when applied to the small dataset of BUS images. Although this study is not an application of a CNN for BUS classification, the results indicate that morphological and textural components of lesions could be valuable additions to a fusion input image of a CNN.

Another interesting discovery from Magnuska et al. suggests that bounding boxes around lesions as CNN input performs just as well as manually segmented masks (2022). Manual segmentation of BUS images to create ground truth masks by radiologists is time-consuming and limits the amount of data available. The study showed that an alternative input of the bounding box to a fully segmented mask performed just as well, with a p-value of 0.071. However, this was performed on a small dataset of just over 500 images total, so verification of the results is needed. Nonetheless, it provides support for the merits of further experimentation on feeding simple inputs into a breast cancer lesion classifying model.

Lastly, a study by Ferreira et al. presents a comparison of input images and CNNs to predict the malignancy of BUS images (2023). The various inputs presented in Table 1 were tested against the following transfer learning models: GoogLeNet, InceptionV3, ResNet, DenseNet, MobileNetV2, and EfficientNet. A visual representation of these inputs is shown in Figure 1. The study resembles the type of comparative analysis that this current research initiative of the BUS Project aimed to conduct but with different variations of inputs.

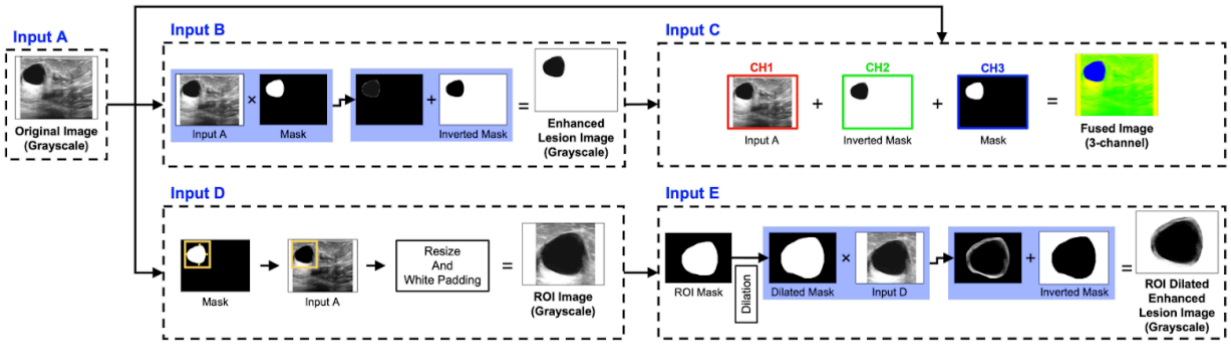
Table 1

Summary of Input Data Variations (Ferreira et al., 2023)

Input	Description
A	Original image
B	Original image with background removed and changed to white
C	Fusion image with Inputs A, B, and the mask image in 3-channels
D	ROI Image of dilated bounding box calculated from mask, cropped
E	ROI dilated enhanced lesion image

Figure 1

Visual Representation of Various Data Inputs (Ferreira et al., 2023)



Conclusion

From the review of existing literature, the opportunity presented by investigating novel inputs and their effects on the performance of CNNs for BUS classification was apparent. Although CNNs have been used with transfer learning methodologies and varying inputs for years now, there is a lack of truly viable CAD applications for BUS lesion classification. The literature suggests that fusion input images into these CNNs increase the model's overall performance, but the quantifiable gain that each type of channel combination provides is unclear. This study aims to fill the gap in existing literature of identifying which types of input images with various localization information included will increase a deep learning model's performance. While some studies have identified options for potential custom input images, this study compares the results of input images with the highest performance across the literature directly. With the identification of promising frameworks and image inputs gathered from past research, combined with existing tools to improve model explainability, the design of this study was then strategically formulated. In the following section, the data collection and processing tools and techniques are discussed.

Chapter 3: Data Collection and Methodology

Introduction

As the overall goal of the BUS Project as a whole is to ultimately create a framework that may be developed into a CAD tool for radiologists to utilize in the hospital setting, the selection of the optimal input image for a deep learning classifier model is a crucial component to explore. To address the question of how varying the input image to include localization information for a BUS deep learning classifier model affects its performance, careful consideration was given to the methodology of this research project. The literature regarding existing related research served as a guide in the formation of the research design, alongside the practical limitations that restrict deep learning applications of CAD breast cancer diagnosis.

Research Design

To address the research question, an experimental and quantitative approach was taken through this study. First, because a binary classification model for predicting whether a tumor was benign or malignant was needed, the top two deep learning frameworks in Python were considered: Tensorflow and PyTorch. Although both open-source frameworks allow for efficient training of deep learning networks, PyTorch was selected because of its ease of use and flexibility in experimentation (Alvi, 2024). The initial network architectures selected reflect the popular and successful networks in the literature, EfficientNet and ResNet. Ferreira et al. found that EfficientNet showed the highest performance when comparing five state-of-the-art models for BUS classification with varying image inputs (2023). Various other researchers have found highest accuracy in BUS classification using ResNet-18, so these two networks were chosen as the basis for creating the PyTorch classifier model (Tsang, 2022).

To evaluate the effect that localization information has on the constructed deep learning models, various input images were created to feed to the models. The raw ultrasound images were transformed using OpenCV tools to produce the contestant input images for comparison (see Appendix C). Inspiration for several of the specific modified input images to consider, shown in Table 2, was drawn from Ferreira et al. (2023), seen in the enhanced lesion inputs with the background removed, as well as Zhang, E. et al. (2020), seen in the BI-RADS feature maps for Inputs F and G. The metrics to evaluate the models' abilities to classify the BUS images were gathered using TorchMetrics, an API offered by PyTorch with over 100 model analysis metrics available. Standard metrics were chosen and collected for evaluating an image classification model including accuracy, precision, recall, and area under the receiver operating characteristic (AUROC).

Table 2

Candidate Image Inputs to PyTorch Model (Baggett, 2024)

Input	Description
A: Original Image	The original image with white padding added to the image, and black padding to the mask. The aspect ratio of the original image is maintained, with an option to resize the image and mask.
B: Enhanced lesion image	The image background was removed using Input A and the segmentation mask, leaving only the lesion visible. The black background was changed to white to differentiate from the lesion.
C: Fused image	Input A, Input B, and the mask were combined into a 3-channel image.
D: ROI image	From the segmentation mask, the bounding box of each lesion is computed and dilated by 30 pixels from each side. Then, the image is cropped using the bounding box, followed by rescaling and padding to the original size.
E: ROI dilated enhanced lesion image	The segmentation mask is dilated by a fixed percentage of its initial area. The computed mask is then replaced by the content of Input D, and the black background is changed to white, creating an image of the lesion with a small border.
F: BI-RADS feature map	The birads-feature map as described by Zhang, E. (2020). Cropped the region to 40 pixels larger than the lesion, resize, and pad as in Input D.
G: Boundary-preserved BI-RADS feature map	This has the same boundary feature map in Input F on the exterior of the lesion boundary, but keeps the original image inside the lesion boundary. This is similar to Input E, but preserves the boundary shape better than the dilation operations used to create Input E.

Deep learning models are data-hungry and can require greater computational power than is available on most CPUs. Due to this reason, a free GPU offered by Google CoLab was utilized to load the images and train the deep learning models initially. As a need for greater computing power arose from incorporating more data and training for a greater number of epochs,

Lightning.AI Studio was used for its varying GPU options. Popular data analysis and visualization Python packages were integral tools in this study as well, including pandas, numpy, scikit-learn, and matplotlib. An additional resource was the legacy code developed by former UW-La Crosse student, Andrei (2022). This legacy code assisted in processing previously collected datasets in a variety of formats and locations into one compact data frame to use for model training. The efforts from previous BUS Project studies were built upon and expanded, including the public and private datasets used to train the model.

Population and Sample

High-quality, expertly annotated BUS datasets are difficult to acquire, but necessary to have to train a deep learning model. For this study, one private dataset provided by the Mayo Clinic and five public datasets were combined for a total of 3,784 images. The data sources are provided in descending quality in Table 3. The quality of each dataset was determined by the confidence that the BUS Project research leads had assigned based on an exploratory review of each dataset.

Table 3

Data Sources in Order of Descending Quality

Dataset Name	Number of Images			Citation
	Benign	**Malignant**	**Total**	
Mayo_Dataset	144	131	275	(Mayo Clinic, private)
BUSBRA	1,268	607	1,875	(Gómez-Flores et al., 2023)
BrEaSt_USG	154	98	252	(Pawłowska et al., 2024)
BUSIS	256	306	562	(Zhang, Y. et al., 2022)
BUS_B	53	110	163	(Yap et al., 2017)
BUSI	446	211	657	(Al-Dhabyani et al., 2020)
Total	2,321	1,463	**3,784**	

The small private dataset provided by Mayo Clinic was deemed the highest in quality with reference to the image and label quality. The size of this dataset was not large enough to train a deep learning model alone, so the additional five public datasets were added to improve the results. To reduce the effect that the quality of one dataset may have on how the model was trained, all image paths were stored in a dataframe, shuffled, and then split randomly into training, validation, and testing sets.

Bias Handling

If only one dataset were to be used for training a model, it would be subject to labeling bias. Part of the motivation for CAD software to assist in the diagnosis of BUS images is due to the variation in interpretations from radiologists for which characteristics of a lesion suggest malignancy. With this in mind, a dataset such as the private Mayo Clinic one provided, when used on its own, would be subject to the bias of the radiologist's discretion when applying the

binary mask annotations. All images used in this study were confirmed with a biopsy to determine the true malignancy label, but the masks that serve as a ground truth for training the models are inherently biased. To mitigate the effects that labeling bias may have on the deep learning models trained, different datasets were combined with a variety of expert-labeled masks. As seen in other studies, mixup augmentations were also used to lessen the impact that the quality of labeling has on the model (Arazo et al., 2020).

Another common source of bias in artificial intelligence within healthcare applications is a lack of diversity of the training data. If a machine learning algorithm is trained on data that lacks diversity, it will be biased towards certain populations (Colón-Rodríguez, 2023). The datasets used in this study were anonymized to protect patient privacy, so the distributions of demographic information for the data were not available to the BUS Project team for review. However, the current literature does not appear to report a link between demographic information and breast ultrasound images. Most bias in breast cancer diagnoses happens from healthcare provider interactions and unconscious biases, such as inattention blindness or dismissing medical concerns for certain demographics (Lamb et al., 2020). Because the malignancy labels for this study were obtained through a biopsy and not a provider's opinion, the unknown diversity in training data was not a concern in addressing biases.

Data Collection and Analysis

Prior to this study, the BUS Project team had acquired 3 public datasets to supplement the private one provided by the Mayo Clinic: BUSIS, BUS_B, and BUSI. From Table 3, these are the lowest quality and only amount to 1,382 images. For this project, the addition of two higher quality datasets, BUSBRA and BrEaSt_USG, more than doubled the size of the training data. All directories to each dataset's grayscale BUS images and binary mask images were first gathered

into a pandas data frame, along with the label that was mapped from the dataset's annotations CSV file. Rows in the data frame corresponding to a single image were randomly split into training, validation, and test sets in a 70/15/15 ratio respectively.

To artificially expand the size of the training set for better results, augmentations to the images were performed. First, the specific transformations to apply as in many image classification studies, geometric augmentations were introduced to the image sets: horizontal flips, translational movements, and slight rotations (Kriti et al., 2020). A random subset of a larger collection of typical augmentations were used for each epoch, as was shown effective in a study by Tupper & Gagné (2024). Vertical flips and large rotations were not used, since BUS images are acquired in a specific orientation based on human anatomy. While training the model, mixup was applied to create combinations of images and their corresponding masks and labels to add more variety to the images and prevent overfitting to specific images (Monigatti, 2023). Normalization was also applied to the grayscale images, whose pixel values originally fell between 0 and 255, to instead fall between 0 and 1. This ensured that the training images and their associated masks were on the same scale, as the masks were binary.

To load the images, a custom Python class was created to load in the data from the dataframe directories, perform the augmentation transformations specified, and return a PyTorch tensor. This custom data loader would call helper functions written by Dr. Jeffrey Baggett to perform the specific manipulations to create the specified input image from Table 2 depending on the input parameters. A custom classifier model class was also constructed that would utilize either a Resnet-18, EfficientNetB0, or EfficientNetB3 model with pre-trained weights and biases as a starting network. Additional networks and ensembling methods could have been added to

increase the predictive power of the model, but these were not implemented for this study as the objective was simply to identify the localization effects of the inputs.

After the data loader and classifier models were defined, the model was trained using a PyTorch interface, PyTorch Lightning. This interface handles the implementation training the deep learning model in a consistent way while still accepting the necessary parameters and allowing for customization of the network. The number of epochs, custom data loader for the training and validation datasets, pre-trained network, loss function, additional transformations, and metrics to gather were passed into the PyTorch Lightning model training function. To evaluate the effects that localization information from different input images has on the model's performance, each of the three models were trained for 50 epochs for each image input type using a cross-entropy loss function, which is standard for classification problems between 0 and 1. Once the models were trained, their performances were evaluated with the test data that had been untouched during the training process. The results from both the model training and testing gathered by PyTorch Lighting were then visualized to assist with interpretation.

Conclusion

While investigating the effects of localization information into a BUS classifier model, the research methodology for this study was developed from an understanding of the current literature and aims to address the gaps identified. While several studies explored the effects of enhanced input images, BI-RADS feature map images, and fusion images as inputs separately and found an increase in performance, no study has compared all of them together to identify an optimal input. The variety of input images, both explored previously and novel combinations, presented in this study directly contrasted against each other gives deeper insight into this topic. A variety of datasets were incorporated to accomplish the project objectives with steps taken to

minimize the effects of labeling bias. The full code outlining the data processing methodology is available for review (see Appendix B). The evaluation metrics, accuracy, precision, recall, and AUROC, were gathered using the PyTorch Lightning interface and presented and discussed in the following chapter.

Chapter 4: Results

Introduction

After the data was acquired and the PyTorch models were constructed and trained for 50 epochs on each type of input image selected for evaluation, several metrics were gathered on the performance of the data on the reserved test set. These metrics are presented in this chapter in the context of the research hypothesis. From the literature review, the first hypothesis was that the best model performance would be produced by Input E, the ROI-enhanced lesion image (Ferreira et al., 2023) or Input G, the BI-RADS feature map image (Zhang, E., 2020). The next highest performance was predicted to be Input C, the fusion image, based on its slightly lower performance when compared to an ROI-enhanced lesion input (Ferreira et al., 2023). All of these inputs were predicted to result in higher performance than Input A, the original ultrasound image used as a control. The second hypothesis was that between the two PyTorch models considered, the EfficientNetB3 one would overall perform better than the ResNet18 one, but both would show similar trends between the input image performances.

Precision, recall, and F1 scores are presented for each class for each input type. Precision is the proportion of true positives out of all predicted positives, with a high precision value for the malignant class indicating that a predicted positive by the model is most likely to be that diagnosis. Recall is the proportion of true positives detected by the model out of all actual positives, meaning that a high recall indicates the model's ability to correctly identify malignant cases as the correct diagnosis. Precision and recall are arguably the most important metrics when considering a CAD tool for breast cancer, as early detection of a positive case is key to a patient's survival rate (Ginsburg et al., 2021). F1 scores represent a balanced combination of both precision and recall, with a high F1 score indicating strong predictive ability.

In addition to precision, recall, and F1 score, the accuracy is also presented for each model. Accuracy is a commonly utilized metric to assess a classification model's performance, but in the case of an imbalanced dataset, less weight is given to accuracy values. Out of all 3,784 images in the combined dataset, only about 38% are malignant. A model that naively predicted benign most of the time could potentially do fairly well. For this reason, recall and F1 scores are the primary metric to review when determining the performances of the Pytorch models.

Research Findings

Once the models were trained and tested on the test data, a classification report was generated using the predicted and actual malignancy labels. A summary of these results are presented in Table 4 and Table 5, displaying the results for ResNet18 and EfficientNetB3, respectively.

Table 4

Summary of Classification Report Results for ResNet18 Model

Input Image Type	Class	Precision	Recall	F1 Score	Accuracy
A: Original	Benign	0.75	0.91	0.82	0.75
	Malignant	0.75	0.46	0.57	
C: Fusion	Benign	0.78	**0.96**	0.86	0.80
	Malignant	**0.87**	0.52	0.65	
E: ROI-Enhanced	Benign	0.85	0.89	**0.87**	**0.83**
	Malignant	0.79	0.73	**0.76**	
G: BI-RADS Feature Map	Benign	**0.88**	0.79	0.83	0.80
	Malignant	0.68	**0.80**	0.74	

When inspecting the results of the ResNet18 model in Table 4, no one model stood out as a clear top performer across all metrics, with each input reporting high variation for precision, recall, F1 scores for both classes. The greatest number for each metric and each class is bolded, showing that each of the three models with localization had at least one of the highest metrics. However, Input E notably resulted in the highest F1 score for the malignant and benign classes and overall accuracy. When considering the hypothesis that either Inputs E or G would produce the best results, it appears that the results are in agreement with this idea. However, it cannot be stated that Input C was significantly worse in its performance than Input G. Surprisingly, the highest value out of the table was for the recall of Input C, suggesting that this fusion input is the strongest when the primary goal of the classifier model is to correctly identify the malignant cases as their true diagnosis. Input A did not have the highest value for any metric, which also supports the hypothesis.

Table 5

Summary of Classification Report Results for EfficientNetB3 Model

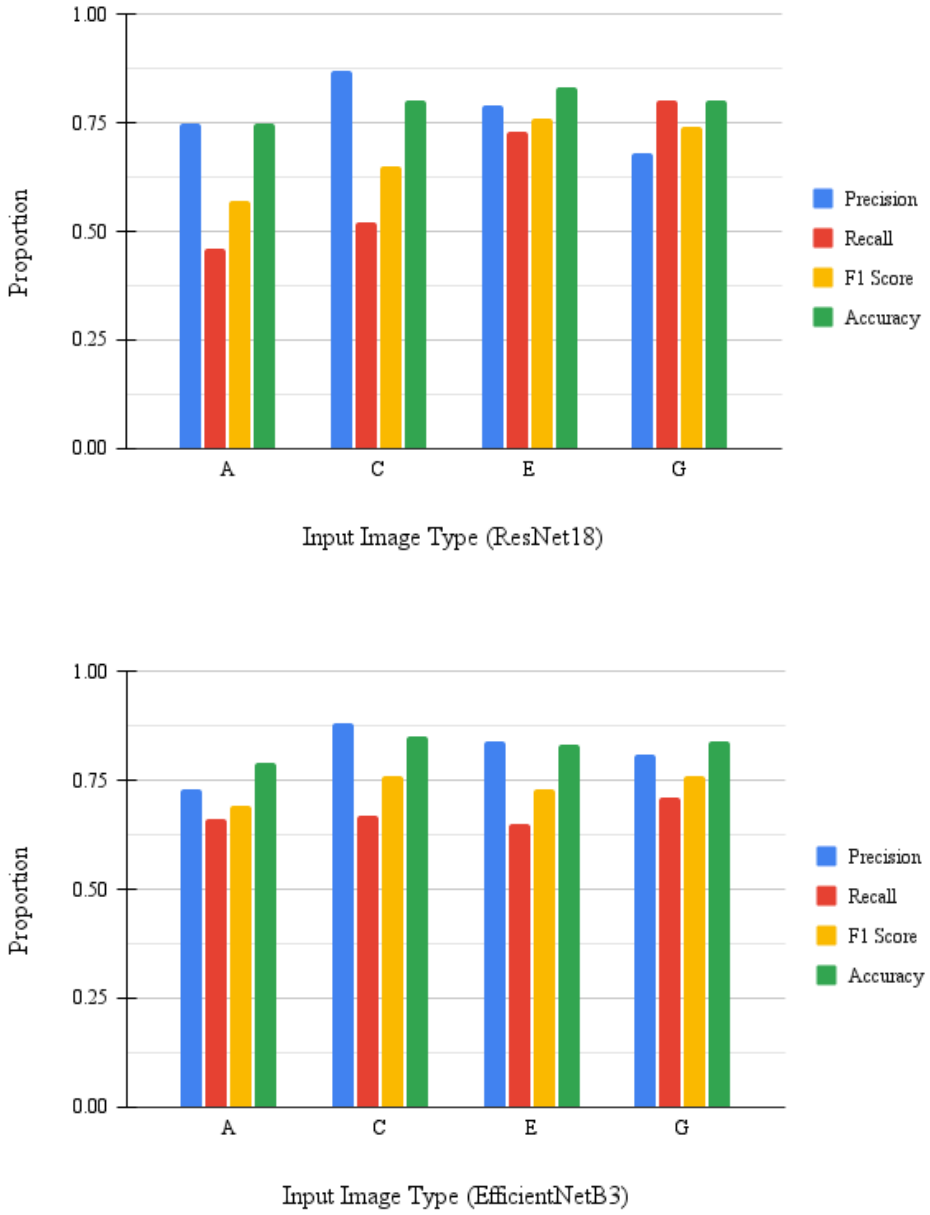
Input Image Type	Class	Precision	Recall	F1 Score	Accuracy
A: Original	Benign	0.82	0.86	0.84	0.79
	Malignant	0.73	0.66	0.69	
C: Fusion	Benign	0.84	**0.95**	**0.89**	**0.85**
	Malignant	**0.88**	0.67	**0.76**	
E: ROI-Enhanced	Benign	0.83	0.93	0.88	0.83
	Malignant	0.84	0.65	0.73	
G: BI-RADS Feature Map	Benign	**0.85**	0.91	0.88	0.84
	Malignant	0.81	**0.71**	**0.76**	

When evaluating the results of the EfficientNetB3 model shown in Table 5, a different trend is seen from the ResNet18 model results. After testing the model with the various inputs, Input C appears to have the strongest performance. Input C showed the highest metrics for F1 score, malignant precision, overall accuracy, and tied for the highest with Input G for benign recall. This was contrary to the initial hypothesis that Inputs E and G would result in the best performance. Input G was the next best performing input, showing the highest metrics for benign precision and malignant recall, tying for the top for malignant F1 score. Unlike the ResNet18 model, Input E did not hold the highest value for any metric. The hypothesis was supported that the addition of localization information would increase the model's performance, as seen in the higher metrics of Inputs C, E, and G over Input A.

To visualize the performance metrics of each input type for identifying the cases of malignant lesions, bar graphs for each model are provided in Figure 2. From these plots, it is evident that for ResNet18, Inputs E and G seem to perform better overall than Inputs A and C. This is not the case for EfficientNetB3, where determining one model's performance over another's is more difficult.

Figure 2

Bar Graphs of Evaluation Metrics for Malignant Class by Model



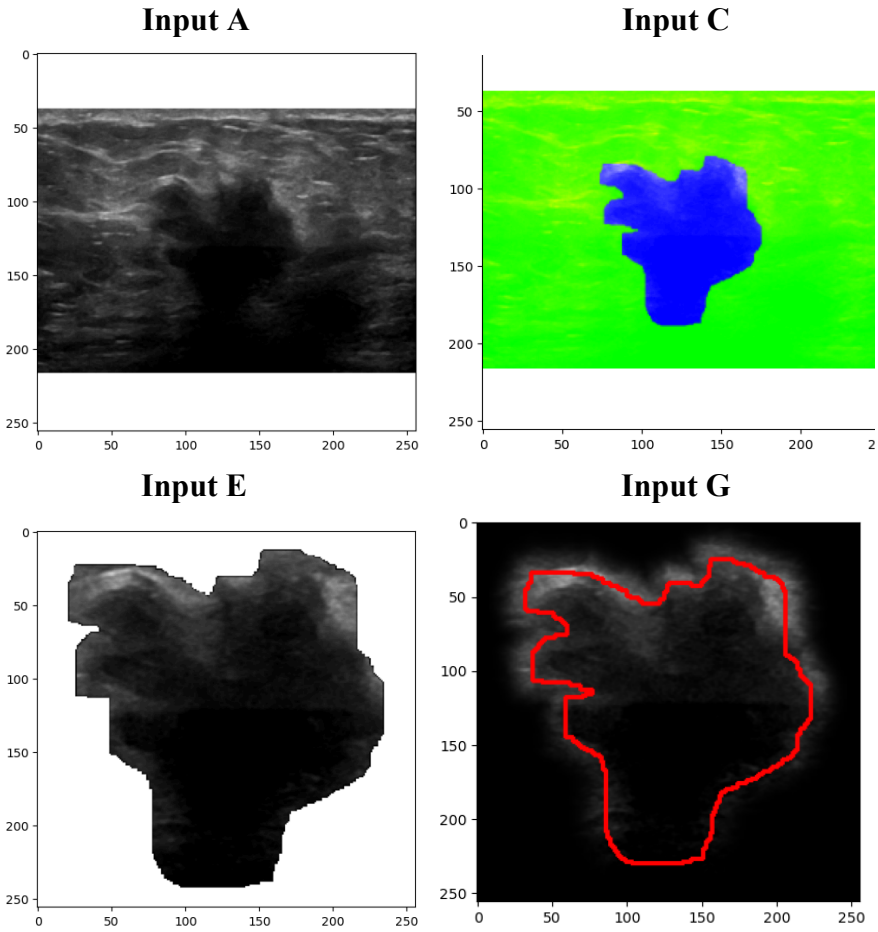
Note. Results for ResNet18 (*above*) and results for EfficientNetB3 (*below*).

Example images are presented for each of the inputs selected to test in Figure 3. Input A was the original BUS image with padding; Input C was the 3-channel fusion image with the

channels being Input A, the enhanced image, and the mask; Input E was the ROI-enhanced image, where ROI was enlarged and the background was removed to increase contrast; and Input G was the BI-RADS feature map.

Figure 3

Example Images of Inputs Selected for Testing



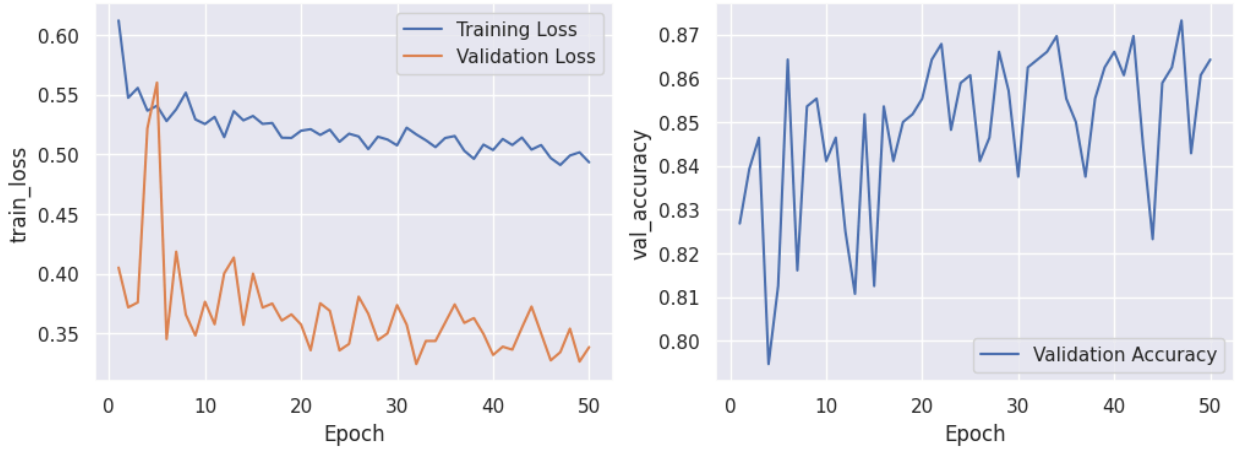
Note: Input descriptions. A: Original image with padding; C: 3-channel (Input A, enhanced image, and mask) fusion image; E: ROI-enhanced, G: BI-RADS Feature Map.

For each input type, a graph was produced of the training and validation loss per epoch, as well as the validation accuracy per epoch. The results for Input E are displayed in Figure 4.

The plots for the other three input types for both models show the same trends, with slightly different levels of noise and values. In general, the training and validation loss trend downwards over the epochs as expected, and the validation accuracy trends upwards. It may be noted that towards the end of the training, it appears that the training and validation loss may still be trending downwards, indicating the model has not fully converged. Although the model may not have fully converged, each model and input was trained for the same number of epochs. The results of their predictive performance are still comparable after 50 epochs of training.

Figure 4

Input E Training and Validation Loss per Epoch (left), Validation Accuracy per Epoch (right)



Another notable feature of Figure 4 is the level of noise present in the validation accuracy. The general trend over the epochs is that the validation accuracy is increasing, which is expected, but there is large variation. This is likely due to mixup augmentations, which combine features and labels to increase the size of the training set (Monigatti, 2023). Although the validation accuracy appears noisy, this augmentation should allow for better generalization on new images and positively affect the model's performance.

Lastly, the hypothesis that the EfficientNetB3 model would perform better than the ResNet18 one is supported in accuracy, but there are still some metrics that are better from the ResNet model. Accuracy in the case of BUS diagnosis is not necessarily the most important metric, so it is difficult to determine that one model is necessarily better than the other. The implications of the models' performances are discussed in the next chapter.

Conclusion

Each of the input types explored increased the model's performance in identifying malignancy when compared to the original BUS image, supporting the initial hypothesis. The evaluation metrics considered when evaluation performance were precision, recall, F1 score, and accuracy. No one input stood out as a clear best candidate to improve the model's predictive performance, especially when taking into account the results of both ResNet18 and EfficientNetB3 models. For ResNet18, Input E showed both the highest F1 score and accuracy, but was outperformed by both Inputs C and G for the EfficientNetB3 model. A detailed discussion on the interpretation and implications of these results in the context of the research hypothesis is presented in the following chapter, as well as limitations of the study.

Chapter 5: Discussion

Introduction

In this chapter, the results of the research study are analyzed in the context of the overall research question. The purpose of this study was to explore the effects that adding localization information to input images of a deep learning classifier model have on its predictive performance, with the long-term goal of a CAD tool in mind. Creating a model that can predict malignancy from BUS images would not only improve the patient experience and increase likelihood of survival, but also present an opportunity for hospitals to save on the expenses associated with misdiagnoses. The results after building and testing the deep learning models with various inputs presented in the previous chapter are summarized and their implications are then discussed in the following sections.

Summary of Findings

After constructing the two PyTorch models and training them for 50 epochs, the reserved testing set was used to evaluate the model performance on new images. For each model, ResNet18 and EfficientNetB3, the metrics for malignant and benign classes of precision, recall, F1 score, and overall accuracy were collected. The results of each model indicated that overall performance was enhanced by the addition of some type of localization information to the classifier model, supporting the hypothesis that the additional information would improve the predictive power. However, the way that performance varied for each input between the two models tested was different than hypothesized. For the ResNet18 model, Input E appeared to be the highest performer, but followed closely by Input G. This matched the hypothesis that either of these two would be the best inputs based on the literature review, with Input C following next.

For EfficientNetB3, Input E performed worse than either Inputs C or G. Input C outperformed Input G, showing the highest of several metrics, including F1 score, malignant precision, benign recall, and overall accuracy. This difference between how the inputs performed between the two models was in conflict with the hypothesis that the trends of how each input performed relative to one another would be similar between models. EfficientNetB3 did show higher accuracy values as expected, but it could not be concluded that the performance in the context of BUS classification for a CAD tool was necessarily better than ResNet18. This mix of performance between models and inputs presents interesting implications for the study and the BUS Project as a whole.

Implications of Findings

As the study sought to identify how localization information in input images affects the deep learning classifier models for BUS diagnosis, the findings imply that exploring localization inputs is well worth the time and effort. For both models, the original BUS image, Input A, demonstrated lower predictive power when compared to the inputs with localization information. The realization that custom input images to increase the performance of a deep learning classifier model is important in the context of the BUS Project, as the optimal input for the model is an important component of the CAD tool. This observation held true for both models, which was expected based on the literature review.

For the hypothesis that Inputs E or G would result in the highest performance, then followed by Input C, the results were mixed between models. For the ResNet18 model, the hypothesis was supported by the findings, but for the EfficientNetB3 model, the opposite was true with Input C performing the best out of the three custom inputs. This was surprising based on the literature that the hypothesis was formed around, as previous studies by Fereirra et al.

demonstrated that an input with an ROI-dilated enhanced lesion was the most effective at improving performance across a variety of models, including a ResNet and EfficientNet model (2023). The differences seen in this study could be attributed to a variety of factors, but the most probable reasons are differences in the size of the training sets as well as the number of epochs used to train. For this study, a total of 3,784 images were combined from five datasets and used, while the Ferreira et al. study used only 733 images from two datasets. This study also only trained for 50 epochs and may not have fully converged, while Ferreira et al. trained for 5,000 epochs. These factors, along with differences in augmentations and specific architectures used, are likely the cause of these unexpected results.

Another hypothesis that was not supported was that the two models would show similar trends, with EfficientNetB3 having better metrics. This hypothesis was also formed from the results of the same Ferreira et al. study as the previous, which indicated that the EfficientNet model showed the strongest predictive power over four other popular pre-trained networks, including a ResNet one. EfficientNetB3 is a more complex network that has a deeper and wider architecture than ResNet18, which was predicted to allow the model to perform better on the custom input images. EfficientNetB3 did have an equal or higher accuracy for each input than its ResNet18 counterpart, but did not show significant improvement. The result that neither model was clearly more powerful than another in terms of predicting the malignancy of custom BUS input images indicates that utilizing some kind of ensembling method may be a good method to improve the robustness of predictions.

Overall, it is difficult to determine exactly what aspects of the input images and architectural framework influenced the predictive power of each model the most. The results imply that the gap in the literature regarding direct comparisons of input images including a

variety of localization techniques should be further explored to fully understand what factors are the most significant in contributing to the outputs of BUS classifier models. Investigating and identifying these factors would enhance the performance of the models while also allowing heightened visibility into the neural network decision-making process.

Limitations of the Study

One limitation of the study that could have influenced the results is the limited access to computational resources. Training each model for each input for just 50 epochs took a significant amount of computational power, which was only accessible towards the end of the research period. Similar studies within the literature have trained models for tens to hundreds of times longer, which may have contributed to better evaluation metrics. From inspecting the training graphs presented in Figure 4, it does look like the model was not fully converged by the 50th epoch, so training the model for even another 25-50 epochs may have improved the performance of each model. Excessive training is not necessary, especially since overfitting would become an issue after the model had converged.

Another limitation is the size of the training set. Although the number of images from the combined dataset was enough to train the model to obtain reasonable predictions, incorporating more high quality datasets would improve the predictive power. These kinds of datasets are difficult to acquire, but it would be good to verify how incorporating more data impacts the results.

Conclusion

The results from training both models on the various inputs showed that localization information improves the predictive power of the models tested. Even though no one model or input stood out as a top contender when trying to optimize the model's performance on

predicting malignancy of BUS images, each showed promise in doing so and suggests that further research is warranted. In the following and final chapter, future recommendations, next steps for the BUS Project, and a conclusion for this study are presented.

Chapter 6: Future Recommendations, Next Steps, and Conclusion

Introduction

Throughout this paper, the topic of adding localization information to an input for a BUS classifier model was explored. The overall goal of this study was to see whether localization information would improve the predictive power of the deep learning models when compared with the standard BUS image and to try to identify the optimal custom input image. After reviewing the results, it is clear that localization does seem to improve the performance of the models, but an optimal input was difficult to determine from the final evaluation metrics. As this study is part of an ongoing collaboration with the University of Wisconsin-La Crosse and Mayo Clinic, future recommendations and next steps to take for the BUS Project are provided in this chapter.

Future Recommendations

Based on the results of this study, further research on understanding and identifying which aspects of localization information are the most important in a model's predictive performance should be conducted. In order to create a CAD tool that consistently performs well and improves both the patient and the provider's experiences in the hospital setting, an optimal image input should be identified. The three options tested in this study, Inputs C, E, and G, showed promise as contestants, but further investigation needs to be conducted. Other candidates could be tested from Table 2, other studies within the BUS literature, or novel input images with localization.

To help determine new input images to test, previous research done by the BUS Project could be utilized. One previous study investigated the use of class activation maps (CAMs) and saliency maps to visualize the regions that are most important to a deep learning model in

making its prediction (Bodart, 2022). These types of tools could be used to brainstorm new inputs to test to improve results further, with the added benefit of enhanced explainability when evaluating the results.

Another recommendation for future research is to add ensembling to the prediction process. The results of this study suggest that different architectural models have advantages and disadvantages for varying input, so feeding an input image to several different models and determining a prediction via majority vote would increase the overall robustness and performance of the process. This would require more computing power to initially train the models, but making predictions is far less computationally expensive. If including ensembling seems to improve the robustness of the predictions, it would be worth the investment done through the initial training. Ensembling is a promising area that should be explored to potentially incorporate into a final CAD tool.

One last recommendation is to incorporate further statistical analyses into the evaluation process to quantify the significance of results. Although classification reports created after generating predictions using the trained models on the reserved test data provide insight into the models' performances, further statistical analysis would be beneficial in understanding the study's implications. This could include providing confidence intervals for metrics or running ANOVA to determine if metrics between inputs or models are significantly different. Adding in this information will help in assessing what the results of a study suggest and making future improvements on other research projects as well.

Next Steps for BUS Project

From the discussion of the results in the previous chapter, a natural next step for the BUS Project to take is to ensure access to adequate computational resources to train the deep learning

models until convergence. The training graphs suggest that running the models from somewhere between 75-100 epochs would likely allow the models to converge, although this would need to be tested systematically. Another step to take based on the results seen in this study is to incorporate more high quality images with associated masks to increase the size of the training data. This may require keeping a close watch on research publications to see when new data sources become available, or tasking a trusted radiologist with annotating more anonymized ultrasound images.

In addition to these steps, the future recommendations provided in this chapter should be evaluated to translate into actionable items. Including tools like CAMs or saliency maps to brainstorm novel input images to test could be explored, as well as searching for other ideas or combinations to try from the existing literature. Ensembling techniques are another option to pursue as a next step, permitted that the computational resources are available to include additional models. Lastly, statistical analysis should be incorporated to improve the interpretability and significance of results after experimentation, to make informed decisions when deciding which techniques and architectures to choose for a final CAD tool.

Conclusion

BUS images present a non-invasive method for early breast cancer detection, but the issue of variability in radiologist determinations of malignancy make it a less reliable method than it could be. Early detection is key to increasing the chances of a patient's survival, which makes creating a CAD tool to improve the accuracy and reliability in predictions a lucrative opportunity that the BUS Project has opted to tackle. Developing this CAD tool using a deep learning framework and custom input images is beneficial to the patient's experience because of

the increased chances of early detection, but also to the hospital by saving billions of dollars associated with misdiagnoses of cancer each year.

In this study, the overall research questions posed at the beginning that guided the experimental design were:

1. How does varying localization information added to an input image affect the model's predictive performance?
2. Does varying localization information of an input image affect different model architectures in different ways?

These questions and a review of the literature lead to the methodology that included constructing two pre-trained PyTorch models and three custom input images, training each combination for 50 epochs, and evaluating on a test set to review the results. The results of the study suggest that adding in localization information to the input of a deep learning BUS classifier model does improve its predictive performance. Exactly which components of the localization information for each input type were the most influential in malignancy determination as well as which one would perform the best long-term could not be concluded. Contrary to the initial hypothesis, the different input images did not show the same trends for each model tested. The ResNet18 model showed that Input E and G were the highest performing inputs, whereas the EfficientNetB3 model showed that Input C was the highest performing one.

From the evaluation of results, future recommendations and next steps for the BUS Project were provided. Further research should be performed to understand and identify how the different localization techniques affect the model's performance and identify other potential custom inputs to optimize the performance even further. Increasing the size of the dataset, ensembling, and more statistical analysis on the results are suggested to drive the BUS Project

towards their ultimate goal of creating a robust CAD tool for BUS diagnosis. This study addressed the gap in the literature regarding a direct comparison of strong candidate input images including fusion images, ROI enhanced lesion images, and BI-RADS feature map images. It accomplished the task of assessing whether incorporating a custom input image with localization information is a worthwhile investment for the research team, as well as providing ideas for further investigation and expansion of the topic.

References

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

Alvi, F. (2024). PyTorch vs TensorFlow in 2024: A Comparative Guide of AI Frameworks. *OpenCV*. https://opencv.org/blog/PyTorch-vs-tensorflow/

American Cancer Society. (n.d.). *Breast Ultrasound*. Retrieved February 12, 2024, from https://www.cancer.org/cancer/types/breast-cancer/screening-tests-and-early-detection/breast-ultrasound.html

Andrei, F. (2022). Semantic Segmentation for Medical Ultrasound Imaging. *University of Wisonsin-La Crosse*. https://github.com/FlorinAndrei/datascience_capstone_project

Arazo, E., Ortego, D., Albert, P., O'Connor, N.E., & McGuinness, K. (2020). Pseudo-Labeling and Confirmation Bias in Deep Semi-Supervised Learning. *2020 International Joint Conference on Neural Networks (IJCNN)*, 1-8. https://doi.org/10.1109/IJCNN48605.2020.9207304.

Baggett, J. (2024). Mask and Image Demo. *University of Wisconsin-La Crosse*. https://datascienceuwl.github.io/CADBUSI/

Bahl, M. (2022). Updates in Artificial Intelligence for Breast Imaging. *Seminars in Roentgenology*, *57*, 160-167. https://doi.org/10.1053/j.ro.2021.12.005

Bodart, Teresa. (2022). Computer-Aided Diagnosis (CAD) of Breast Cancer: Methods of Model Explainability. *University of Wisconsin-La Crosse*. https://datascienceuwl.github.io/CADBUSI/documents/Capstones/Computer-Aided%20Diagnosis%20(CAD)%20of%20Breast%20Cancer-%20Methods%20of%20Model%20Explainability%20-%20Teresa%20Bodart.pdf

Colón-Rodríguez, C.J. (2023). Shedding Light on Healthcare Algorithmic and Artificial Intelligence Bias. *U.S. Department of Health and Human Services*. https://minorityhealth.hhs.gov/news/shedding-light-healthcare-algorithmic-and-artificial-intelligence-bias#:~:text=Healthcare%20algorithms%20and%20AI%20bias,used%20to%20train%20computer%20programs

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

Ciria-Suarez, L., Jiménez-Fonseca, P., Palacín-Lois, M., Antoñanzas-Basa, M., Fernández-Montes, A., Manzano-Fernández, A., Castelo, B., Asensio-Martínez, E., Hernando-Polo, S., & Calderon, C. (2021). Breast cancer patient experiences through a journey map: A qualitative study. *PLoS One*, *16*(9). https://doi.org/10.1371/journal.pone.0257680

Dai, J., Lei, S., Dong, L., Lin, X., Zhang, H., Sun, D., & Yuan, K. (2021). More Practical AI Solution: Breast Ultrasound Diagnosis Using Multi-AI Model Ensemble System. *Tsinghua University*. https://arxiv.org/ftp/arxiv/papers/2101/2101.02639.pdf

Daoud, M.I., Bdair, T.M., Al-Najar, M., & Alazrai, R. (2016). A Fusion-Based Approach for Breast Ultrasound Image Classification Using Multiple-ROI Texture and Morphological Analyses. *Computational and Mathematical Methods in Medicine*, *Vol. 2016*(6740956). http://dx.doi.org/10.1155/2016/6740956

Ferreira, M.R., Torres, H.R., Oliveira, B., de Araujo, A.R.V.F., Morais, P., Novais, P., & Vilaca,

J.L. (2023). Deep Learning Networks for Breast Lesion Classification in Ultrasound Images: A Comparative Study. *Institute of Electrical and Electronics Engineers*. https://doi.org/ 10.1109/EMBC40787.2023.10340293

Ginsburg, O., Yip, C.H., Brooks, A., Cabanes, A., Caleffi, M., Dunstan, J.A., Gyawali, B., McCormack, V., McLaughlin de Anderson, M., Mehrotra, R., Mohar, A., Murillo, R., Pace, L.E., Paskett, E.D., Romanoff, A., Rositch, A.F., Scheel, J.R., Schneidman, M., Unger-Saldaña, K., … Anderson, B.O. (2021). Breast cancer early detection: A phased approach to implementation. *Cancer*, *126*(10), 2379-2393. https://doi.org/10.1002%2Fcncr.32887

Gómez-Flores, W., Gregorio-Calas M.J., & Coelho de Albuquerque Pereira, W. (2023). BUS-BRA: A breast ultrasound dataset for assessing computer-aided diagnosis systems [dataset]. *Medical Physics*. https://doi.org/10.1002/mp.16812

Hao, K. (2019). We analyzed 16,625 papers to figure out where AI is headed next. *MIT Technology Review*. https://www.technologyreview.com/2019/01/25/1436/we-analyzed-16625-papers-to-figure-out-where-ai-is-headed-next/

Kriti, Virmani, J., & Agarwal, R. (2020). Deep feature extraction and classification of breast ultrasound images. *Multimedia Tools and Applications*. https://doi.org/10.1007/s11042-020-09337-z

Lamb, L.R., Fonseca, M.M., & Verma, R. (2020). Missed Breast Cancer: Effects of Subconscious Bias and Lesion Characteristics. *RadioGraphics*, *40*, 941–960. https://doi.org/10.1148/rg.2020190090.

Lo, J., Cardinell, J., Costanzo, A., & Sussman, D. (2021). Medical Augmentation (Med-Aug) for

Optimal Data Augmentation in Medical Deep Learning Networks. *Sensors*, *21*, 7018. https://doi.org/10.3390/s21217018

Magnuska, Z.A., Theek, B., Darguzyte, M., Palmowski, M., Stickeler, E., Schulz, V., & Kießling, F. (2022). Influence of the Computer-Aided Decision Support System Design on Ultrasound-Based Breast Cancer Classification. *Cancers*, *2022*(14), 27. https://doi.org/10.3390/cancers14020277

Monigatti, L. (2023). Cutout, Mixup, and Cutmix: Implementing Modern Image Augmentations in PyTorch. *Medium*. https://towardsdatascience.com/cutout-mixup-and-cutmix-implementing-modern-image-augmentations-in-PyTorch-a9d7db3074ad

National Breast Cancer Foundation. (2024, January 17). Breast Cancer Facts & Stats. Medically reviewed on June 15, 2023, by Lillie D. Shockney. https://www.nationalbreastcancer.org/breast-cancer-facts/

Ong, M.S. & Mandl, K.D. (2015). National expenditure for false-positive mammograms and breast cancer overdiagnoses estimated at $4 billion a year. *Health Affairs (Millwood)*, *34*(4), 576-83. https://doi.org/10.1377/hlthaff.2014.1087

Pawłowska, A., Ćwierz-Pieńkowska, A., Domalik, A., Jaguś, D., Kasprzak, P., Matkowski, R., Fura, Ł., Nowicki, A., & Zolek, N. (2024). A Curated Benchmark Dataset for Ultrasound Based Breast Lesion Analysis (Breast-Lesions-USG) (Version 1) [dataset]. *The Cancer Imaging Archive*. https://doi.org/10.7937/9WKK-Q141

Ragb, H., Ali, R., Jera, E., & Buaossa, N. (2021). Convolutional neural network based on transfer learning for breast cancer screening. *Cornell University*. https://doi.org/10.48550/arXiv.2112.11629

Rajput, D., Wang, W.J., & Chen, C.C. (2023). Evaluation of a decided sample size in machine

learning applications. *BMC Bioinformatics*, *24*, 48. https://doi.org/10.1186/s12859-023-05156-9

Rawat, W. & Wang, Z. (2017). Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review. *Neural Computation*, *29*, 2352-2449. https://doi.org/10.1162/NECO_a_00990

Tsang, S.H. (2022). Brief Review — Breast Ultrasound Image Classification and Segmentation Using Convolutional Neural Networks. *Medium*. https://sh-tsang.medium.com/brief-review-breast-ultrasound-image-classification-and-segmentation-using-convolutional-neural-c20112cabd5e

Tupper, A., & Gagné, C. (2024). Analyzing Data Augmentation for Medical Images: A Case Study in Ultrasound Images. *arXiv*, 2403.09828. https://doi.org/10.48550/arXiv.2403.09828

Yap, M.H., Goyal, M., Osman, F., Marti, R., Denton, E., Juette, A., & Zwiggelaar, R. (2020). Breast ultrasound region of interest detection and lesion localisation. *Artificial Intelligence in Medicine*, *107*(2020), 101880. https://doi.org/10.1016/j.artmed.2020.101880

Yap, M.H., Pons, G., Marti, J., Ganau, S., Sentis, M., Zwiggelaar, R., Davison, A.K., & Marti, R. (2017). Automated Breast Ultrasound Lesions Detection using Convolutional Neural Networks [dataset]. *IEEE Journal of Biomedical and Health Informatics*. https://doi.org/10.1109/JBHI.2017.2731873

Zhang, B., Vakanski, A., & Xian, M. (2021). BI-RADS-NET: An explainable multitask learning approach for cancer diagnosis in breast ultrasound images. *Institute of Electrical and Electronics Engineers*. https://doi.org/10.1109/mlsp52302.2021.9596314.

Zhang, E., Seiler, S., Chen, M., Lu, W., & Gu, X. (2020). BIRADS features-oriented semi-supervised deep learning for breast ultrasound computer-aided diagnosis. *Physics in Medicine and Biology*, *65*, 125005. https://doi.org/10.1088/1361-6560/ab7e7d

Zhang, Y., Xian, M., Cheng, H.D., Shareef, B., Ding, J., Xu, F., Huang, K., Zhang, B., Ning, C., & Wang, Y. (2022). BUSIS: A Benchmark for Breast Ultrasound Image Segmentation. *Healthcare*, *10*(4), 729. https://doi.org/10.3390/healthcare10040729

Appendix A

BI-RADS Scores and High-Level Definitions

BI-RADS Score	High-Level Definition
0	*Incomplete* - Insufficient information
1	*Negative* - Nothing new or abnormal found
2	*Benign* - Non-cancerous lesion
3	*Probably Benign* - Follow-up within a short time frame
4	*Suspicious* - Biopsy should be considered
5	*Highly Suggestive of Malignancy* - Biopsy strongly recommended
6	*Already Confirmed Malignancy* - Used for treatment response logging

Appendix B

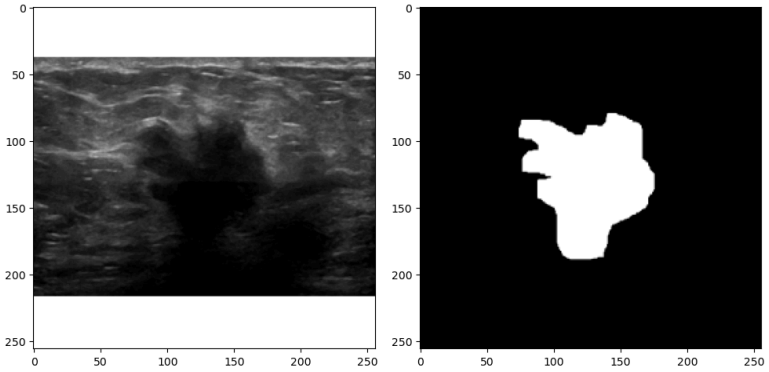
GitHub Repository for Source Code

https://github.com/rachelkim97/UW-Lax-Capstone

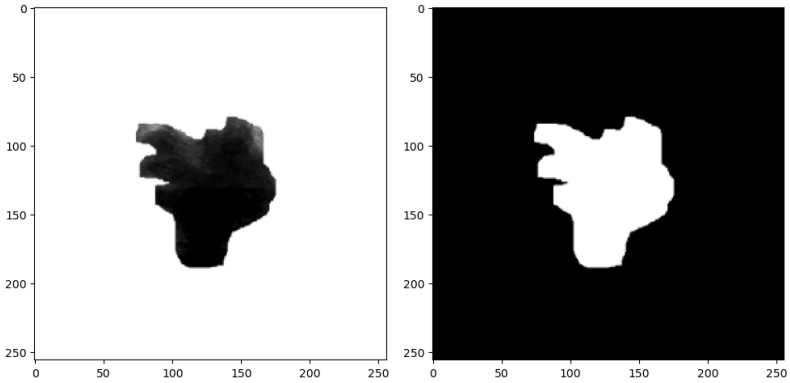
Appendix C

Proposed Input Image Examples with Masks (Baggett, 2024).

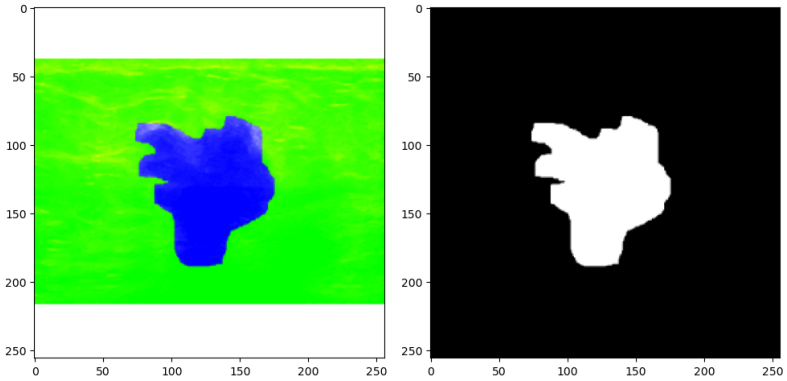
Input A



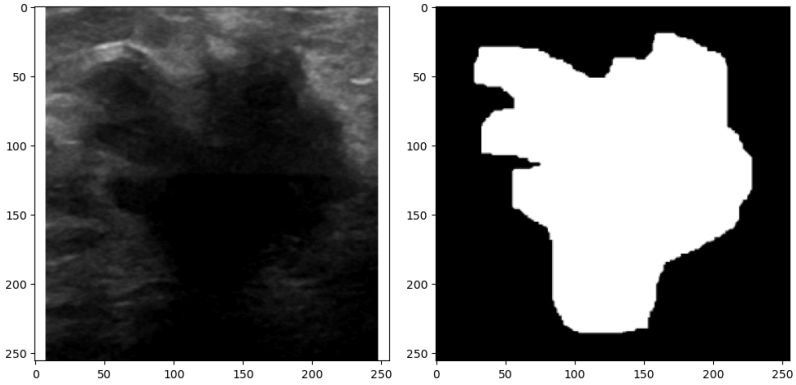
Input B



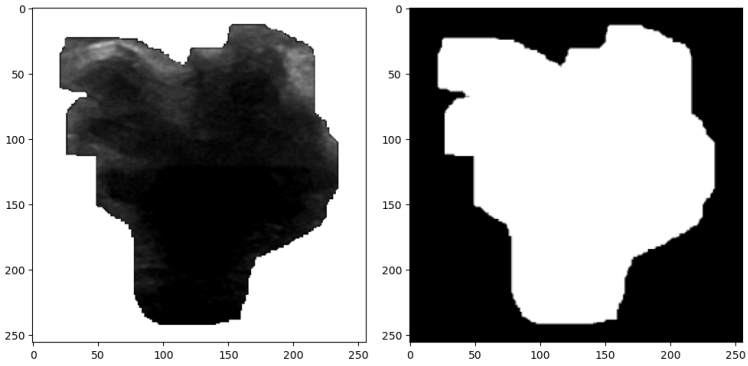
Input C



Input D



Input E



Input F (left) and G (right)