GT Summer 2023 Practicum
Final Report

AI+ First Technology: Breast Cancer Ultrasound Diagnosis -- Team 1

Tyler J Lang, tylerjlang@gatech.edu

Brian Nutwell, bnutwell3@gatech.edu

Feng (Frank) Wen, fwen9@gatech.edu

Abstract – In this project we develop an ML pipeline to analyze new ultrasound images, locate any lesions, and correctly predict whether each lesion is malignant or benign, in order to support physician effectiveness when performing breast cancer screening.

1 INTRODUCTION

Breast cancer is a leading cause of female mortality, and since 2020 it has been the most commonly diagnosed form of cancer worldwide. In China in particular, more than 300,000 new cases were diagnosed and 70,000 deaths reported from breast cancer in 2015, and the numbers continue to increase (Gu et al., 2022). Put another way, roughly 1 in 500 Chinese women are newly diagnosed with breast cancer every year despite quite limited testing coverage.

While mammography is the preferred method of breast cancer diagnosis in the USA and other western nations, it is considered less effective in China. As summarized by Gu et al.:

> …Mammography-based breast cancer screening is not very practical in China … because Chinese women tend to have small and dense breasts, and because the peak age of breast cancer onset is younger than that in Western women, both of which are known to affect the diagnostic accuracy and effectiveness of mammography.

Due to these factors, ultrasound imaging (US) is often preferred for diagnosis in China. But analyzing US images of breast lesions typically requires a high degree of expertise. As with other diseases, a lack of testing or a mistaken benign diagnosis can lead to delayed treatment and negative health outcomes, while an incorrect malignant diagnosis can cause unnecessary follow-up testing as well as anxiety for the patient.

A further complicating factor is that despite rising cancer incidence rates and recent government efforts to increase testing coverage, most Chinese women still do not receive regular breast cancer screening; as of 2019 only 30.9% of women aged 35-64 in the

country had ever been screened for breast cancer, with significant disparities across the economic spectrum (Zhang et al., 2023). So Chinese medical authorities are challenged to dramatically increase the number of screenings performed each year while maintaining confidence in diagnoses which typically require a high level of human expertise; this challenge is especially acute in rural and poorer areas where such expertise, medical facilities, and other resources are at a premium.

One possible solution is the implementation of deep learning models which can quickly evaluate ultrasound images and support physicians in making a diagnosis – either reinforcing the judgment of inexperienced clinicians, or allowing the operator to evaluate more patients in a given period of time while ensuring quality of the results. Current advancements in CNN architectures for image recognition, such as YOLO, ResNet and DenseNet pre-trained models, along with the availability of sophisticated training and deployment libraries like PyTorch, make such a task quite accessible. This is the project approach suggested by AI+ First Technology Co. Ltd., which we will explore in this report.

2 PROJECT DESCRIPTION, STRUCTURE OF THIS REPORT

For this project, the sponsor (AI+ First Technologies) provided a training dataset of labeled Breast Ultrasound images collected from within their network of hospitals (see examples in *Figure 1*). The team was asked to develop a machine-learning algorithm and pipeline to accurately diagnose the presence of breast cancer in the images. This report will detail our project activities, technical approach, and results.

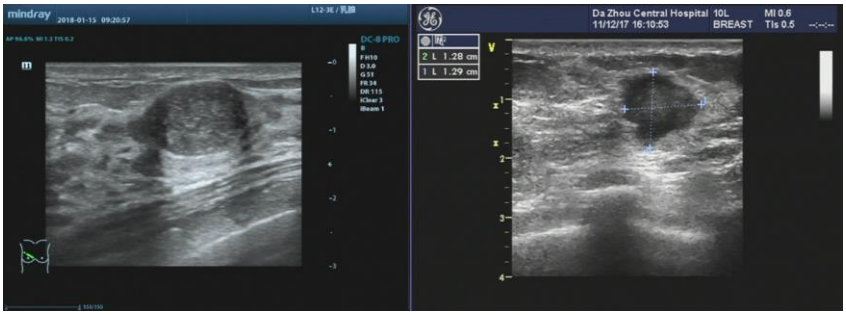

Figure 1 - Sample images for benign (left) and malignant (right) tumors

In section **3 Review of Existing Practices** we examine existing literature in the domain to identify best practices. And we further collect input from a physician working in this space as a Subject Matter Expert, to help align our tool's output with user needs.

Section **4 Data Sources** explores the sponsor-provided dataset and additional publicly-sourced datasets, with discussion of data cleaning, image enhancement, and image augmentation techniques. Section **5 Proposed Approach** outlines the main technical elements used and the structure of our test/train pipelines. **Section 6** details the performance trials, selection and final tuning of the ML algorithms used. **Section 7** summarizes the best observed performance from our modeling activities, while the remainder of the report deals with suggested next steps (**Section 8**) and summary/conclusions (**Section 9**). **Section 10** explains team member contributions, and **Section 11** is the **Appendix** including References and further details of various elements.

3 REVIEW OF EXISTING PRACTICES

Commented [FW1]: I feel like this section is better to be Section 2, before the Data Source.

Commented [BN2R1]: Sure, that's reasonable. Think I started with a chronological approach (got the problem, then researched related stuff) but putting this earlier makes sense.

Commented [BN3R1]: Moved it here to Section 3 after the project description

3.1 Literature Review

The team reviewed existing published work related to ML diagnosis of breast ultrasound images. A summary of the most relevant recent papers and observations are shown in *Figure 2* below. From this review there were a number of key conclusions:

- There appear to be **two dominant approaches** to using ML tools for this problem. Roughly half the papers (including Eroğlu et al., 2021; Wang & Yao, 2022; and Meraj et al., 2021) took a **sequential two-step approach**, in which one model is trained for segmentation, to identify a region of the image which contains a tumor, and a second model is trained to classify the resulting segmented image as benign or malignant. Other papers (such as Sahu et al., 2023) utilized an **object-detection approach**, in which one model was trained to perform both tasks in a single analytical step. In addition, several papers focused only on optimizing one individual step in the process, such as segmentation (Byra et al., 2020; Gu et al., 2022; Thomas et al., 2023) or image enhancement (Singh et al., 2020).
- Papers used a **wide variety** of existing pre-trained **ML models** (such as ResNet, U-Net, and even non-CNN classifiers like SVM or KNN) or proprietary custom-built CNNs. There was little agreement about tuning parameters or any "best" approach, but nearly all papers mention that **ImageNet was used to pre-train** the neural networks for image recognition before final layers were re-trained on the problem dataset.

- Although a few papers sourced their own images, several **public datasets** of pre-cleaned and labeled breast ultrasound images exist. These can be used to augment the limited training image sets in our project. Examples of these datasets include OASBUD and BUSI – further discussion is included in **Section 4 Data Sources**.
- Though there was little consistency in the metrics reported, best-in-class results appear to be a **DSC** (Dice Similarity Coefficient) **score ~0.9 for segmentation, and accuracy of 0.95 or higher for final classification**.
- Nearly every paper mentioned significant image-preprocessing steps, often performed manually, before training or validating the data models. This requirement would present a limitation in deploying an ML model for real-time field use by working physicians. Our team will attempt to automate this step as much as possible.

Based on the literature survey, we will pursue a number of the suggested best practices: perform CLAHE + interpolation for image pre-processing, utilize public datasets to augment our sponsor images, employ pre-trained image recognition models, and use DSC / mAP metrics for segmentation and classification respectively. In addition we will investigate both 2-step and 1-step modeling approaches.

Lead Author	Year	Datasets	Humans	Image Pre	Segmentation	Seg. Metric	Classification	Class. Metrics	Notes
Gu	2022	14K images, proprietary	Experts + non-experts	Grayscale, bounding box	Manually by experts, lesion + 50-pixel border	N/A	VGG-19, pretrained	AUC = 0.908	Collected their own images, demonstrated performance vs. experts, demonstrated ability to improve non-expert diagnoses
Meraj	2023	B2 (train), B+O (test)	N/A	Resize, CLAHE, user in the loop	U-Net + Otsu quantization to improve as needed	(None reported)	Bagged-boost (and others), 5/10-fold CV, salience, feature regularization	Accuracy 93%	Quite recent paper. Didn't use CNN for classification, interesting. Segmentation method didn't seem automated.
Wang	2022	B1, B2, B3	N/A	CLAHE + anisotropic diffusion for noise	Original R2AttU-Net, segmentation only used to enhance image	IoU	FCOS anchor-free detection network (mmdetection, ResNet50)	mAP = 0.902	Enhance image with one CNN, don't segment, then classify with another CNN that works well on differently-scaled images
Thomas	2023	B2, O, R, U	N/A	Not discussed	Benchmarked several: Mask R-CNN was best	DSC = 0.851	N/A	N/A	Very recent paper with good analysis of various architectures on common datasets: focused on segmentation only
Byra	2020	B2, O, U	N/A	Remove annotation, resize, 3x3 median filter	SK-U-Net, SK blocks, custom cost function = 1-Dice	DSC = 0.900	N/A	N/A	Highly cited paper, good technical discussion of innovation in CNN
Sahu	2022	B2	N/A	None, input BUSI directly to the model	N/A		Hybrid model, ShuffleNet + ResNet	Accuracy 98%	Studied many variations of hybrid model using MRI and US images. No separate classifier model, just directly predicting 2 or 3 classes.
Erdoglu	2021	780 images, proprietary	N/A	Augment data by copy+ rotate, 500x500	Hybrid Alex + ResNet + MobileNet, concatenate & mRMR	None: used layer outputs as features	Other classifiers KNN, SVM	Accuracy 95.6%	Trained 3 CNN models, extracted hidden layer outputs as features, then performed classification on a mRMR-reduced feature-set.
Singh	2019	Ultrasound videos	4 experts	CLAHE + CNN					Paper only studied enhancement of images (videos). Used a CNN for despeckling, and ran experiments for best CLAHE settings.

Figure 2 -- Summary of literature survey observations

3.1.1 *Metrics*

To evaluate the performance of each model, we will pursue 3 key metrics suggested by the literature.

DSC (Dice Similarity Coefficient) is a measure of congruence of two shapes: in our case, the expert-provided bounding box label around a tumor vs. the ML-predicted bounding box from a segmentation model. DSC is calculated as $DSC(A,B)= 2(A \cap B)/(A+B)$ where $\cap$ is the intersection of the two areas and (A+B) is the summed areas.

mAP (mean Average Precision) is a measure of prediction quality across multiple classes and confidence levels. It computes the average Precision (AP) over all classes, where Precision is the ratio of true positives to the sum of true and false positives. The AP for each class is calculated as the area under the Precision-Recall curve, providing a single-figure measure of quality across recall levels. The two most commonly used mAP metrics for object detection are mAP50 and Map50-95:

- mAP50 refers to mAP evaluated at an Intersection over Union (IoU) threshold of 0.50. IoU measures the overlap between two bounding boxes (the predicted box and the ground truth). An IoU of 0.50 means that the detection is considered "correct" if the overlap of the ground truth and prediction boxes is 50% or more.
- mAP50-95 means calculating the mAP at IoU thresholds from 0.50 to 0.95 with a step size of 0.05 (i.e., 0.50, 0.55, 0.60, ..., 0.95). This provides a more comprehensive evaluation of the model's robustness across different overlap requirements, reflecting how precise the object detection model is at varying degrees of strictness.

In addition, we will examine Recall as a measure of how well our models correctly identify the true-positive (malignant) tumors – as the risk and potential health impact of incorrect cancern diagnosis is asymmetrical, we may wish to avoid false-negative diagnosis at the expense of some more false-positives.

3.2 Subject Matter Expert Interview

To augment our understanding from the literature review and sponsor explanation, the team prepared a questionnaire for a working ultrasonographer in the sponsor's network and target region (rural China hospitals). The full questionnaire and response are included in the **Appendix**.

The expert clarified that a physician will personally conduct the ultrasound (including multiple views of the same patient) and then examine the images to make a diagnosis, during a single working session. Implications for an ML tool are that it would be most helpful to provide a rapid suggested diagnosis including confidence level, so that the physician could decide on the spot whether to collect additional images. The physician also mentioned that multiple images and other factors such as age are taken into account when making a diagnosis, along with specialized medical expertise, so the model diagnosis from a single image should be best viewed as a physician support tool and not an output to be shared directly with the patient.

Key insights (as translated by the sponsor) which support this understanding include: "*The most important aspect [when making a diagnosis] is ... observing the morphological features of the image*"; and *"[when developing such a tool] I think it's necessary to have a good understanding of the relevant medical knowledge related to breasts, such as their structure, tissue organization, and lymphatic metastasis.*"

4 DATA SOURCE

For this study we utilized both the sponsor provided data as well as several external datasets of breast ultrasound images which are publicly available.

4.1 Datasets

The sponsor provided a dataset with 1146 breast ultrasound images. The team also identified several public datasets during the literature review process because larger training sets will generally lead to better results in machine learning.

Eventually, the team selected two public datasets: 1. Breast Ultrasound Images Dataset (BUSI); 2. Open Access Series of Breast Ultrasound Data (OASBUD). Due to the nature of the breast ultrasound images, each image only contains one type of tumor (i.e., benign, or malignant).

Table 1 summarizes the information on Datasets used in this project. There is a bit of data imbalance between type benign and malignant (about 3:2).

Table 1 – Summary of Datasets

Datasets	No. of Images	Labels by Tumor Types	
		Benign	Malignant
AI + Training	1149	662	487
BUSI	647	437	210
OASBUD	200	96	104
Total Training	*1996*	*1195*	*801*
AI + Testing	239	132	107

For the training/validation/test split, the team selected 180 images (104 benign, 76 malignant) from the AI+ dataset as the validation set. The left 969 images from the AI+ data and the two public datasets to be used as training data.

One thing to note was that the images provided by the sponsor included metadata about the associated patient ID (i.e., there were cases where several images could be from the same patient). Therefore, for the training/validation split during the training process, the team ensured that no images from the same patient were split across the training and validation dataset to prevent any data leak issue.

For the test set, the team used the entire external test dataset provided by the sponsor.

Commented [BN4]: @Wen, Feng can you add a comment about how to select images (i.e. same patient/case handling)

4.2 Data Cleaning

The team observed that the images often contain information from the operation panels for the AI+ dataset. This information usually includes letters, digits, and shapes different from the targeted ultrasound results and may cause noise which can mislead the models during model training.

Considering ultrasound images taken by different brands of devices have different layouts, the team manually categorized the images based on the devices (e.g., Esaote, GE, Samsung, Philips etc.), then cropped them to preserve only the ultrasound image part, as shown in *Figure 3* below.

Since the two selected public datasets don't have the noise information from operation panels, the above cropping process also brought consistency when mixing data from different datasets.

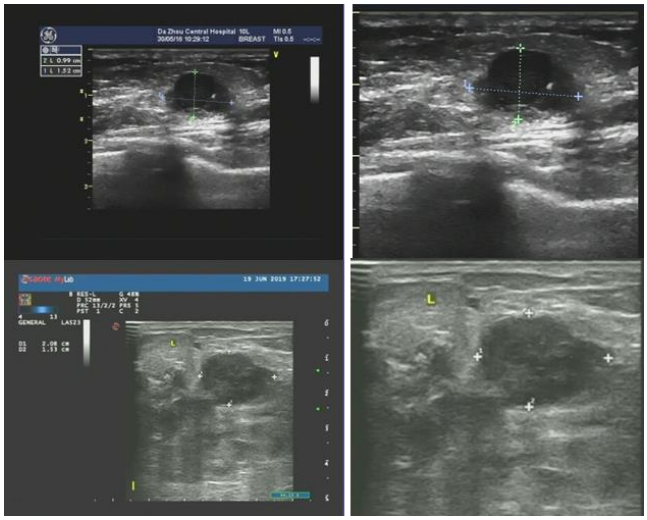

Figure 3 - Sample original (left) and cropped (right) images from sponsor provided dataset

4.3 Image Enhancement

For black-and-white images in the medical industry such as X-ray or ultrasound results, it's a standard process to preprocess the images with image enhancements such as Adaptive Histogram Equalization (AHE) or Contrast Limited Adaptive Histogram Equalization (CLAHE) to increase the contrast and make the features (e.g., contours and textures) of the targets clearer and more stand out.

Based on best practices from the literature survey (Section 3), the team decided to use the CLAHE method for image enhancements. The default CLAHE method uses a combination of Uniform distribution for histogram shape and Bilinear Interpolation for combining the transformation functions computed for neighboring and regions. Singh, 2019 showed an innovative approach of CLAHE using a combination of Rayleigh distribution and Lanczos-3 Interpolation. The team tried both these CLAHE methods in this project (as CLAHE Default and CLAHE Best Practice).

Samples of the above two image enhancements approaches are shown in *Figure 4*. Visual examination suggests that CLAHE Best Practice approach creates images with the clearest contrast.

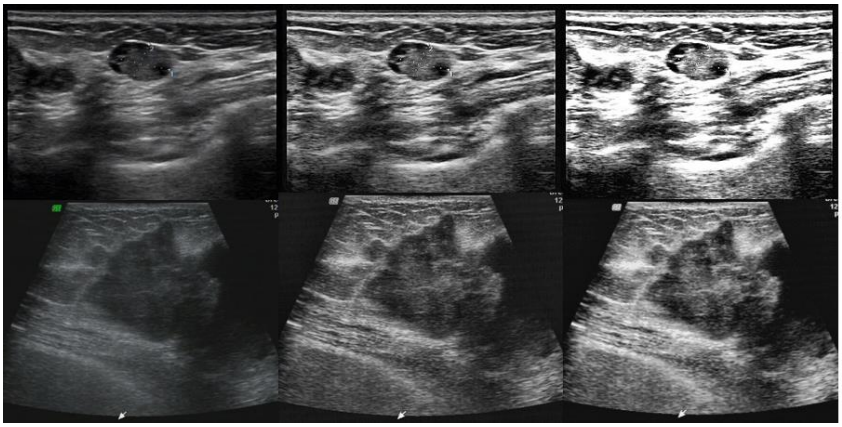

Figure 4 -- Samples of Image Enhancements, from Left to Right:
Original Image; Image Enhancement using CLAHE Default; Image Enhancement using CLAHE Best Practice

4.4 Data Augmentation

Data augmentation is an effective technique to increase training by creating modified copies of the existing data. General data augmentation methods for computer vision include random cropping, random shifting, color distortion, etc.

For this project, the team embedded the data augmentation of training data during the training process using the Python library Albumentations. Various combinations of data augmentation methods have been tested; see section 6.1.4 for further details of our approach.

5 PROPOSED APPROACH

As the "1-step image recognition" and "2-step segmentation + classification" approaches each have potential advantages, we experimented with both.

5.1 "1-Step Image Recognition" Concept

One set of trials involved establishing a single-model pipeline for training a YOLOv8 Object detection model to both segment and classify the tumors. This approach is illustrated in *Figure 5* below.

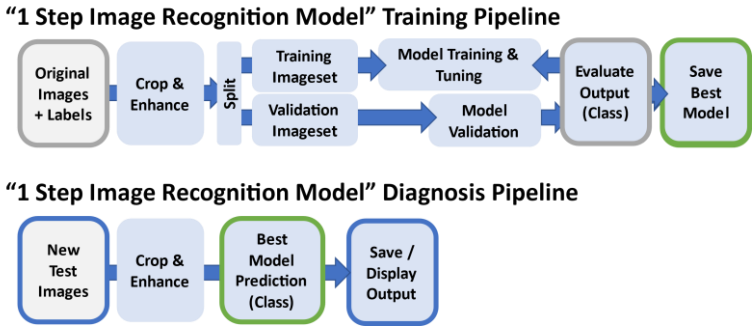

Figure 5 - 1-step pipeline approach to breast ultrasound diagnosis

5.2 "2-Step Segment + Classify" Approach

Due to its excellent performance in object segmentation alone, we further used the YOLO model in our two-step approach as an object segmentation tool. In this set of trials, we trained the YOLO model just to look for tumors and provide their bounding box coordinates, which we then use to crop the images for further classification using a DenseNet model architecture, pre-trained on the popular ImageNet-1k dataset. Given a tumor exists in the image and is correctly detected in our object segmentation model, the classification task then becomes a simple binary task, in which confidence of output can be directly interpreted from the softmax output of prediction, interpreted as a percent-value between 0 and 1 as to how confident the model is in determining the tumor type. This second approach is illustrated in *Figure 6*.

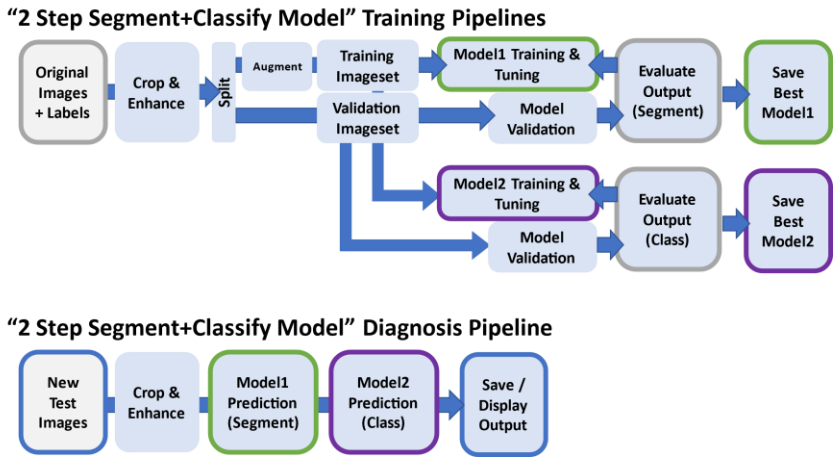

Figure 6 - 2-step pipeline approach for breast ultrasound diagnosis

The "Crop & Enhance" step in both workflows above follows the methods explained in Section 4. Details of model tuning and performance results will be explained in the subsequent sections of this report.

6 MODEL CONSTRUCTION AND TUNING

6.1 "2-Step Segment + Classify" Approach

Following best practices from a variety of literature in the field of deep learning for cancer detection, we built a two-model pipeline, in which one model is used for object detection, which would be used to locate any tumor, and then the location of the tumor would then be used to crop the image which is passed to an image classification model. We evaluated the following model types for the respective tasks. Results from the model trials are explained in subsections 6.1.1 ~ 6.1.5.

- Segmentation: Faster R-CNN, YOLOv8
- Classification: DenseNet121, ResNet50, MobileNet, ensemble CNN methods, KNN, Kernel SVM, and Boosted SVM

6.1.1 *Faster R-CNN for Segmentation*

As a baseline, we used a Faster R-CNN ResNet50 architecture to predict the bounding boxes of the tumors as a first step to detection, which would then be passed to a classification model to decide on the type of tumor. After some basic fine-tuning of the hyperparameters, the model was only able to achieve a DSC score of 0.35 on validation data. While the model was able to create a bounding box around 90% of tumors, it suffered from inexactness, in which multiple bounding boxes would surround the tumor with similar prediction scores, or the predicted area would be too large, providing imprecise cropping for the second-stage classification model. See *Figure 7* for examples.

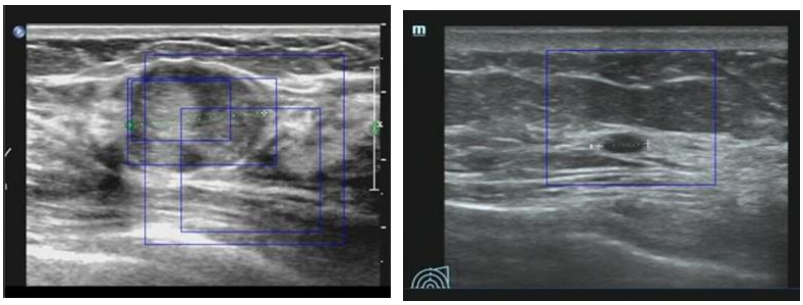

Figure 7 – Sample bounding box output from Faster R-CNN model.
Note multiple bounding boxes (left) and overlarge detection area (right).

While further improvements are certainly possible with fine-tuning, a trial of the YOLO model for the same segmentation task showed much more promising initial results (see the next section). As a result, the team did not pursue the Faster R-CNN model development further.

6.1.2 *YOLOv8 for Segmentation*

We also examined YOLOv8 for the segmentation task – see **Section 6.2** below for discussion of the YOLO model's unique technical approach and advantages.

The YOLO model can be converted to a segmentation model by the simple expedient of labeling all images with bounding boxes and a single class, "tumor." Trials of the available tuning parameters quickly demonstrated good accuracy, as demonstrated by *Table 2* and *Figure 8*.

Table 2 – Summary of YOLOv8 Segmentation Performance on Validation Set

Model	Optimizer	Learning Rate	Batch	Image Size	Total Epochs	Best Epoch	DSC	mAP50	mAP50-95
YOLOv8n	SGD	0.001	16	160	100	81	0.9	0.988	0.644

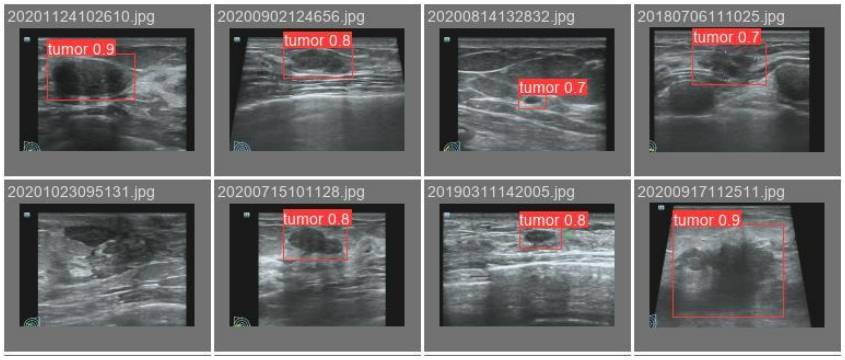

Figure 8 - Predicted bounding boxes from YOLOv8 segmentation model. Note that tumors are generally well identified, with the exception of the image at bottom left in which no bounding box was returned.

We can also see that a small number of images (<5%) were not able to be segmented by this model, returning no bounding boxes at all. In these cases, we decided to pass the entire cropped image to the classification algorithm, but we expect relatively poor diagnosis accuracy for such cases. This is a topic for future study and/or tuning.

6.1.3 *Tumor Classification – non-CNN models*

Commented [BN5]: Brian add this section

For the classification task, we evaluated several non-"deep learning" approaches that have proven powerful for complex pattern recognition in other applications. Approaches studied included KNN, Kernel SVM, and Boosted SVM, which were all mentioned in the literature as possible classifiers.

For each model, hyperparameters were tuned using GridSearch and 5-fold cross validation. Models were trained and tested against 3 variations of the training images:

- Image data resized to 224 x 224 pixel size, comprising roughly 50k features
- Image data reduced using Principal Component Analysis (PCA) to capture the maximum variance in the data, with 128 features
- Image Data reduced using Minimum Redundancy Maximum Relevance (mRMR) to capture the most relevant variance in the data, with 128 features

Accuracy on the validation set was generally poor with all approaches, and average precision did not exceed 0.6 for any case. Significant overfitting was observed with the full image data and the PCA data. However, for all 3 models the reduced mRMR feature-set of data actually produced higher test accuracy than the full image data. It is possible this feature reduction approach would have value for other algorithms; we leave this suggestion for future work. See *Table 3* below.

Table 3 – Classification Performance of non-CNN Algorithms

Model	Dataset	Cropped	Enhanced	Tuning Parameters	Data	Test mAP
KNN	AI+	Y	N	n_neighbors = 61	224x224	0.505
					PCA	0.495
					mRMR	0.571
Kernel SVM	AI+	Y	N	C=1, kernel = 'rbf'	224x224	**0.576**
					PCA	**0.563**
					mRMR	**0.595**
Boosted SVM	AI+	Y	N	C=1, kernel = 'rbf', n_estimators=10, algo=SAMMER	224x224	N/A*
					PCA	0.550
					mRMR	0.578

** – not performed due to performance limitations of Boosting algorithm on 50k features*

Based on the superior performance of the CNN classification models (see the next section), the team did not pursue further tuning of KNN and SVM based classifiers.

6.1.4 *Tumor Classification – CNN models*

Pytorch's Torchvision package contains a variety of developed CNN models trained for classification, each architecture specializing in different methodologies to increase accuracy and/or computational efficiency. As a baseline, three types of models were trained to explore the optimal architecture for this task: DenseNet, MobileNet, and ResNet. All models were loaded with weights pre-trained on the ImageNet-1K dataset, which includes over a million images classifying 1000 objects. While the objects the model was trained on to obtain the initialization weights are not related to or reminiscent of tumors, using such an initialization we presume allows the model to fine-tune much more quickly using the tumor data due to having weights tuned for detecting basic shapes and edges from the objects it was previously trained on. And as expected, the model within just a few epochs has exponential improvements in accuracy, jumping from 50% to 70% or higher

within the first 3-5 epochs. Running these models for 60 or even 100 epochs increased test accuracy even with quickly diminishing marginal improvements. Due to plateauing accuracy around 30 epochs, an exponential learning rate decay of 10% is used to allow the model to further fine-tune for slight increases in model accuracy, albeit causing the model to train more slowly, and thus requiring 60 or more epochs. See *Table 4*.

Table 4 – Summary of CNN Classification Performance on Validation Set

Model	**Opti-mizer**	**Learning Rate**	**LR Decay**	**Weight Decay**	**Total Epochs**	**Best Epoch**	**Accuracy**	**Precision / Recall**	**F1**
Resnet50	Adam	0.0001	10%	0.01	60	52	0.8771	.84/.87	.86
MobileNet	Adam	0.0001	0%	0.005	70	9	0.7989	.77/.84	.78
DenseNet	Adam	0.0005	20%	0.01	60	59	**0.9274**	**.89/.95**	**.92**

Additionally, image enhancement methods were used to express further the tumor and its characteristics, which, in all models, improved accuracy, and thus all model performances shown below were based on models trained using the enhanced images. Image enhancement techniques include first reading the image as a grayscale image, converting it to an array, normalizing the array for values between 0 and 1, using a Rayleigh CLAHE algorithm to equalize the histogram of the pixel values, which increases contrast, helping to make textures and object boundaries more notable, without losing information from over contrasting. Finally, Lanczos interpolation is used to enlarge the image and zoom in on the tumor in a way that smoothly interpolates the values between the pixels across the images.

Once attaining initial baseline accuracies across a variety of model architectures, including ResNet, DenseNet, MobileNet, Inception, and ResNext, among others, ResNet and DenseNet architectures proved to be the most effective, especially at properly detecting malignant tumors, having higher recalls overall than other model architectures. For further testing, we compared various baseline accuracies across a ResNet-50 model and a DenseNet-121 model fine-tuned with various learning rates, learning rate decays, weight decays for normalization, and semi-final layer dropout levels. Learning rates specifically varied from 1e-3 to 5e-5, with 1e-4 and 80-90% decay found to be optimal across both architectures.

Table 5 – Summary of ResNet Classification Performance on Validation Set

Opti-mizer	Learning Rate	LR Decay	Weight Decay	Total Epochs	Best Epoch	Accuracy	Precision / Recall	F1
Adam	0.0001	1%	0.005	15	14	**.8603**	**.89/.76**	**.82**
Adam	0.0001	10%	0.005	10	8	.8547	.84/.81	.82
Adam	0.00005	5%	0	20	12	.8492	.84/.79	.81
Adam	0.0001	1%	0	10	9	.8547	.89/.75	.81
AdamW	0.0001	1%	0	10	9	.8492	.85/.77	.81
AdamW	0.0001	1%	0.005	10	8	.8547	.84/.81	.82

ResNet is a deep architecture that maintains accuracy by using identity mapping, in which skip connections are added to a block of layers to pass the input directly to the output along with the convolved input to prevent accuracy from degrading deeper down in the network, which was a common problem in past deep networks. See *Table 5* for ResNet training performance.

$y = F(x, \{W_i\}) + x$ —— (1) → identity function

Table 6 – Summary of DenseNet Classification Performance on Validation Set

Opti-mizer	Learning Rate	LR Decay	Weight Decay	Total Epochs	Best Epoch	Accuracy	Precision / Recall	F1
SGD	0.0005	20%	0.01	70	69	.8883	.86/.88	.87
AdamW	0.0005	20%	0.01	70	62	.8994	.85/.92	.88
Adam	0.0005	30%	0.01	70	58	.8827	.85/.88	.86
Adam	0.0005	20%	0	70	67	.8939	.88/.87	.87
Adam	0.0005	20%	0.01	70	59	**.9274**	**.89/.95**	**.92**

DenseNet, similarly to ResNet, involves layers receiving previous inputs via shortcuts, but instead of receiving just the previous layer's input, a DenseNet model compiles all previous layers' output, creating a dense block that grows with each layer, at a rate called the growth rate, which is directly proportional to how deep the layer is. To control size

and learn higher-level features, a transition layer pools with a kernel size 2x2, and a stride of 2, while the dense block preserves its growing size. In theory, the continuous addition of past contextual information contained in the dense block is the reason DenseNet models commonly dominate ResNet models, while being more computationally demanding.

Various methods were attempted to improve accuracy even further, but with little success. Ensembling models' predictions together using both hard and soft classifications (adding either the 0/1 predictions or the softmax values between 0 and 1 per model per image, and taking the ceiling of the average) consistently performed worse than our best DenseNet model, due to a handful of certain images in the validation and training sets simply proved too difficult for all models, in which case no model's input helped the best model, and in fact only worked to reduce the efficacy of the best model toward the worst model's performance.

Due to the small dataset consisting of 1149 images, in which 180 are separated for validation, we also used data augmentation methods to add variety to the images across each epoch, in which, prior to being loaded into the model and predicted on, certain alterations via the Albumentation module are performed to make each given image distinct across various epochs. The types of changes performed include shifts, scaling, an rotating the images by a maximum of 5%, random crops, shifts in color by as much as 15 points per R/G/B channel, as much as a 50% increase in brightness, various levels of normalization, and horizontal and vertical flips. This did help increase accuracy consistently by a few percentage points, and thus was adopted in the pipeline for all training tasks. External datasets were also included in the training process to add variety, but did not lead to a trained model that exceeded performance of the best model, which achieved 95% recall on the validation set.

6.1.5 *Classification Pipeline Specifications*

For the classification pipeline, first the data is loaded from json files, and the x1, x2, y1, y2 coordinates are loaded to crop the tumors for training. Data is separated into training and validation sets based on keeping images from the same patient together and balancing a ~ 15% validation set size. Pytorch datasets are created as classes, which define the actions to be performed on the image upon being loaded in a batch via a PyTorch dataloader. In the Dataset class, the image is loaded by name and local path, the bounding box boundaries are loaded and used to crop the image, the image is resized to 160x160 pixels, and the data augmentation functions via Albumentation are applied randomly, at random levels, as described above. The training batch size is set to 32 to maximize accuracy while

maintaining a manageable batch size that does not exceed computational limits. Then, a custom function is used to load a model from TorchVision given the specifications provided, such as the model type, learning rate, weight decay, the learning rate decay, the last layer's dropout percentage, and the optimizer being used. Finally, for training, the model is run per training batch, the optimizer performs backpropogation to tune the weights of each layer, and final validation predictions are performed to report on per-epoch accuracy after all batches in the training set are run.

Sample output is shown in *Figure 9* below.

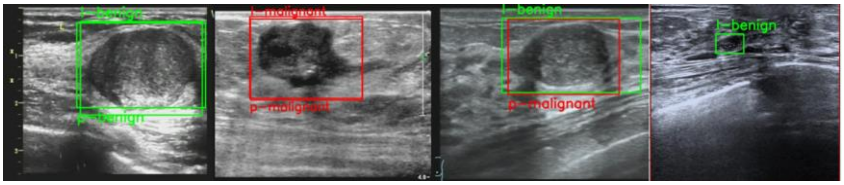

Figure 9 - Example output from trained 2-step pipeline. Ground truth labels are shown with "l-diagnosis" and predicted labels as "p-diagnosis". From L to R: correct benign label; correct malignant label; incorrect label; and undetected tumor with no predicted label. Note relatively good segmentation performance.

6.1.6 *Classification Pipeline Validation*

With the two best models trained for segmentation and classification, a testing pipeline was constructed to first predict a tumor bounding box and then predict a diagnosis on the contents of the segmented area.

Despite the relatively good performance of the two models individually, the stacking effect of sequential models resulting in lower overall performance. See *Table 7* below.

Table 7 – Summary of 2-Step Approach Performance on Validation Set

Model	Segment DSC	Accuracy	Precision	Recall
YOLOv8n (segment) + Densenet121 (classify)	0.88	0.756	0.735	0.658

6.2 "1-step Image Recognition" Approach

For this approach we experimented with the YOLO model architecture, using the Python Ultralytics package. The You Only Look Once (YOLO) model is a profound advancement in object detection algorithms, characterized by its real-time execution speed and

high precision. Unlike traditional two-step object detection methods such as R-CNN and its variants, which first identify regions of interest and then classify these regions, YOLO follows a unified, end-to-end approach. It divides the input image into a grid, each cell predicting multiple bounding boxes and associated class probabilities.

The structural advantage of YOLO resides in its singular convolutional neural network (CNN) that simultaneously predicts multiple bounding boxes and class probabilities for these boxes in one pass—avoiding the "multiple sliding frames" examination step for image sub-regions, a processing-intensive task commonly used by other CNN image processing algorithms. As a result, YOLO dramatically increases efficiency and scalability.

Diverging from other object detectors, YOLO's approach allows it to account for contextual information by treating object detection as a regression problem to spatially separated bounding boxes and associated class probabilities. It encourages holistic properties and enables a generalized model for object detection, translating to enhanced speed and performance. Despite its simplicity and speed, YOLO matches the mean average precision of state-of-the-art systems like Faster R-CNN, making it a potent model for real-time object detection and video applications as well as our ultrasound problem.

For this project, the team selected the most advanced and best-performant model in the YOLO family – YOLOv8. YOLOv8 introduces several enhancements over its predecessors:

- YOLOv8 is an anchor-free model, predicting object centers directly instead of the offset from a known anchor box. This improvement reduces the number of box predictions, thus accelerating the Non-Maximum Suppression process.
- Architectural modifications of YOLOv8 include replacing the initial 6x6 convolution in the stem with a 3x3 convolution, altering the main building block (C2f replaces C3), and changing the conv's kernel size of the bottleneck from 1x1 to 3x3 to make it directly concatenate features, decreasing parameter count and tensor size.
- YOLOv8's training routine focuses on online image augmentation. Notably, it uses mosaic augmentation, amalgamating four images, bolstering the model's learning capacity.

6.2.1 *YOLOv8 Tuning and Results*

The team trained and fine-tuned different YOLOv8 models with different hyperparameters, image enhancement, and cropping methods. *Table 8* below summarizes the best

results for various combinations during the training process (results were tested on the validation set).

- All the models were trained by the training and validation set specified in **Section 4**, using the same training/validation split.
- "Dataset" indicates whether the model was trained based on the provided AI+ dataset or includes the extra public datasets of BUSI and OASBUD.
- "Cropped" indicates whether the images were cropped to remove the information from the operation panel. The images from training, validation, and test set were processed consistently (i.e., all cropped or none cropped), as the team found that inconsistent datasets will lead to bad model performance.
- "En." indicates whether the images were enhanced using the CLAHE Best Practice method. As the team found out CLAHE Best Practice method generally works better than the CLAHE Default method, this is the only image enhancement method used in this stage. Also, consistency was kept for images from the training, validation, and test sets (i.e., all enhanced or none enhanced).
- "Lr" indicates the learning rate used during the training process.

Table 8 – Summary of 1-Step Object Detection Approach Results on Validation Sets

Model	Dataset	Cropped	En.	Lr	Batch Size	Image Size	Optimizer	mAP 50	mAP 50-95
yolov8n	AI+ & Public	Y	N	0.01	32	640	Adam	97.0	63.5
yolov8n	AI+	Y	N	0.01	32	640	Adam	94.5	60.1
yolov8n	AI+ & Public	Y	Y	0.01	32	640	Adam	95.4	60.5
yolov8n	AI+ & Public	Y	N	0.01	32	640	SGD	96.9	61.9
yolov8n	AI+ & Public	Y	Y	0.01	32	640	SGD	95.5	59.4
yolov8s	AI+ & Public	Y	N	0.01	32	640	Adam	96.8	61.1
yolov8s	AI+	Y	N	0.01	32	640	Adam	94.1	59.2
yolov8m	AI+ & Public	Y	N	0.01	32	640	Adam	96.2	62.6
yolov8m	AI+	Y	N	0.01	32	640	Adam	90.1	54.8
yolov8l	AI+ & Public	Y	N	0.00001	16	640	Adam	98.7	66.5
yolov8l	AI+ & Public	N	N	0.00001	16	640	Adam	86.7	58.2
yolov8l	AI+	Y	N	0.00001	16	640	Adam	94.5	60.1
yolov8l	AI+ & Public	Y	Y	0.00001	16	640	Adam	97.2	61.9
yolov8x	AI+ & Public	Y	N	0.00001	16	640	Adam	96.0	65.1

Observations from tests and trials include:

- YOLOv8l has the best performance over the other four models. The reason that YOLOv8x has worse performance could be due to the limitation of a relatively small dataset used for this project.
- The Adam optimizer was slightly better than the SGD optimizer, but the gap is very small.
- Applying CLAHE Image Enhancement didn't seem to improve the model performance.
- Adding public datasets did improve the model performance.
- Cropping the images to remove the operation panel's information helped increase the model performance on the validation sets.

The team eventually decided to use the YOLOv8l model with an input image size of 640 for testing, as this guaranteed the model's performance while making trade-offs between training time and required GPU memories. *Table 9* shows the benchmark results provided by YOLOv8 developers, which indicates that YOLOv8l has a very close performance compared to YOLOv8x and is much more efficient.

Table 9 – Benchmark of YOLOv8 models on COCO val2017

Model	Image Size	mAP50-95	Params (M)	FLOPs(B)
YOLOv8n	640	37.3	3.2	8.7
YOLOv8s	640	44.9	11.2	28.6
YOLOv8m	640	50.2	25.9	78.9
YOLOv8l	640	52.9	43.7	165.2
YOLOv8x	640	53.9	68.2	257.8

6.3 Final Approach(es) chosen

Based on the results above, the team selected a pipeline composed of YOLOv8n segmentation + DenseNet classification for the 2-step approach, and a tuned YOLOv8l model for the 1-step approach. Performance of both approaches on the External Test Set will be reviewed in **Section 7** below.

7 MODEL PERFORMANCE & REFLECTION

7.1 2-Step Segmentation and Classification Approach

Using our best YOLOV8l model for segmentation and best DenseNet model for classification, we constructed a single training pipeline that crops the tumors in the external test sets, and then predicts on tumor classification. Thus, with two models involved, the overall accuracy is dependent on both models, and any upstream error in the YOLO model's cropping exacerbates error in the final classification model, and thus, given our final validation accuracy of 93% with our chosen DenseNet model, this 93% thus serves as a ceiling performance for the model on the external test set, given that any error, even if slight, in the YOLO model will only serve to decrease this accuracy; and also due to standard expectations of decreased performance on a test set which will have unique characteristics unseen during the training and tuning phases of model development.

As an assessment on each individual model's performance with the external test set, we found the YOLO model to produce a DSC score of 0.775, and the DenseNet model to achieve 80% accuracy with 75% recall on images pre-cropped with the tumors' coordinates provided with the test images. When combined, using YOLO-produced bounding boxes to crop the tumors and the DenseNet model to predict on those croppings, our final pipeline was able to achieve 66.1% Accuracy and 55.1% Recall.

These accuracy levels are somewhat worse than observed in training even individually, and the overall combined pipeline has insufficient performance. See *Table 10* below. Further study of possible improvements to the pipeline will continue.

Table 10 – Summary of 2-Step Approach Results on External Test Set

Model	Segment DSC	Accuracy	Precision	Recall	# Undetected
YOLOv8n (segment) + Densenet121 (classify)	0.775	0.661	0.641	0.551	2

7.2 1-Step Object Detection Approach

Table 11 below summarizes the best 5 YOLOV8l models fine-tuned from Section 6.2.1 on the external test sets.

- Column "Model", "Dataset", "Cropped" and "En." Indicates same meaning as described in Section 6.2.1.

- "# Undetected" indicates the number of images that the YOLOv8 model failed to generate any bounding boxes.

Table 11 – Summary of 1-Step Object Detection Approach Results on External Test Set

Model	Dataset	Cropped	Enhanced	mAP50	mAP50-95	# Undetected
yolov8l	AI+ & Public	N	N	87.0	49.0	30
yolov8l	AI+ & Public	N	Y	82.0	46.0	39
yolov8l	AI+	N	N	47.9	21.6	123
yolov8l	AI+ & Public	Y	N	90.6	53.1	21
yolov8l	AI+	Y	N	47.5	23.0	124

Observations from the final test include:

- CLAHE enhancement didn't seem to improve the model performance. Possible causes of this the raw images contain subtle variations. Applying CLAHE might overemphasize certain features while suppressing others. This could cause the model to overfit to the contrast-enhanced training data. In addition, CLAHE may lead to exaggerating noise or minor features in the ultrasound images. These could mislead the model during training, resulting in poorer performance.
- Adding the two public datasets has greatly increased the performance of the multiclassification model. The two public datasets contained new features from the AI+ dataset that could match those in the test dataset.
- Cropping the images to remove the operation panel's information helped increase the model performance on the external test sets. This could be due to reduced noisy data (e.g., texts, letters, etc.), which kept the model focusing on effective features of tumors.

As shown in *Table 11*, the best model achieved mAP50 0.906 and Map50-95 0.531 on the test sets. This performance has reached the same level of YOLOv8 benchmark model on the COCO val2017 dataset (refer to *Table 9*), but with far less training data. The hyperparameter used was: learning rate 0.00001, batch size 16, image size 640, optimizer Adam.

Sample output from the 1-step model is shown in *Figure 10*.

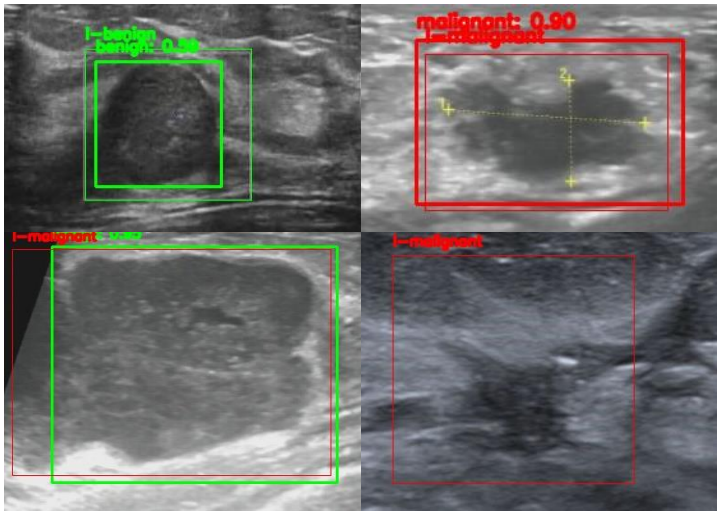

Figure 10 - Example output from trained 1-step pipeline. Ground truth labels are shown with "l-diagnosis" and predicted labels as "p-diagnosis". The upper two images show correct prediction samples. The lower left image shows a wrong prediction sample; the lower right image shows no object detected.

7.3 Discussion

The 1-step YOLOv8 image recognition model has clearly superior performance and should be pursued as the recommended solution. Possible reasons could include:

- **Feature Information Loss in 2-Step Approach**: In the 2-step approach, the image was segmented by the YOLO detector and then passed to a classifier. This segmented image might not contain all the original information (e.g., contextual information) present in the full image. Contextual information, such as the surrounding tumor structures or other global elements, might be vital for correct tumor classification. The classifier might perform worse when we cut this information out in the cropping process.
- **Model Discrepancy**: The 2-step approach used different models (YOLO and then ResNet, DenseNet, etc.), which may cause a discrepancy in feature

extraction. The 1-step approach will learn and extract features more consistently and holistically, which may result in better performance.

- **Optimization Difficulty in 2-Step Approach**: In the 2-step approach, the detector (i.e., yolov8 binary classification model) and classifier (e.g., Densenet) were optimized separately. This means the detector might not be optimized best for the classifier, and vice versa. For instance, the detector could be optimized to localize the tumors perfectly, but the classifier might need more context, as discussed above. In the 1-step approach, the localization and classification are optimized together, which could result in better performance.
- **Accumulation of Variance in 2-Step Approach**: In the 2-step approach, any errors made in the initial detection step could propagate to the classification step. For instance, if the YOLO detector did not perfectly localize a tumor, it could lead to worse classification performance. This is not an issue in the 1-step approach where the detection and classification are done in one step.

8 SUGGESTED NEXT STEPS

Additional suggested investigation items, that could not be completed within the scope of this project, include:

- The provided training dataset, while well organized and labeled, includes a number of graphical artifacts around the tumor areas in many images – in the form of dotted lines or solid cross marks (see *Figure 11*). These appear to be coordinate markings from the physicians labeling suspected lesions; if so, such markings might not be available in the field when this tool is deployed to evaluate unknown images or to support less-expert users. It is possible that the CNN training process is actually training models to recognize these markings, rather than to recognizing the shape of a tumor itself. Presuming these markings will not be present in future unknown images, the team suggests to retrain all models on images without any such line or cross artifacts in the tumor regions.

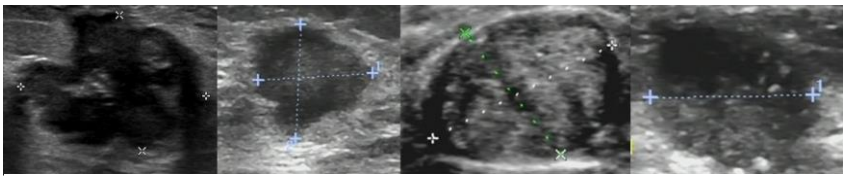

Figure 11 – Example AI+ training images with graphical artifacts included (cropped to show detail).

- Physicians working with cancer imaging use the BI-RADS ranking system to classify the level of cancer risk diagnosed for a patient. This ranking can be used to determine the urgency and appropriate approach for any additional treatments prescribed. Gu et al., 2022, demonstrated an approach to supporting physician BI-RADS ratings with ML model results. The team suggests to utilize our model's prediction confidence and output a suggested BI-RADS rating directly from the tool. This may require further study of the ranking methodology, in co-operation with physicians.
- This study was performed on a small number of AI+ images. The performance of our models greatly increased with the addition of external test sets of images. Expanding the pool of training data from the machines of interest could potentially further improve detection and classification.

9 SUMMARY & CONCLUSIONS

In this project the team evaluated a variety of emerging and best-practice ML approaches to diagnosing breast cancer based on ultrasound images. Both 1-step image-recognition and 2-step segmentation + classification models were explored and tuned. A variety of techniques for image cleanup, enhancement, and augmentation were demonstrated.

It was demonstrated that a tuned YOLOv8l model can perform very well on this task, with mAP50 scores >0.9 and high recall. This model, an example of powerful latest-generation ML tools, is quite simple to train and deploy despite the sophistication of its inner workings.

10 STATEMENT OF TEAM CONTRIBUTIONS

Tyler J. Lang took the lead on training and tuning numerous classification and segmentation algorithms for the 2-step model approach, and contributed to all deliverables.

Brian Nutwell took the lead on constructing the 2-step pipeline and drafting interim project report deliverables, organized the team meetings and Teams/Github sites, and contributed to all deliverables.

Feng (Frank) Wen took the lead on data cleaning and enhancement including public datasets, led the algorithm training and tuning for the 1-step model approach, and contributed to all deliverables.

Each team member contributed an equivalent amount of effort.

11 APPENDIX

11.1 References

Byra, M., Jarosik, P., Szubert, A., Galperin, M., Ojeda-Fournier, H., Olson, L., O'Boyle, M., Comstock, C., & Andre, M. (2020). Breast mass segmentation in ultrasound with selective kernel U-Net convolutional neural network. *Biomedical Signal Processing and Control*, *61*. https://doi.org/10.1016/j.bspc.2020.102027

Eroğlu, Y., Yildirim, M., & Çinar, A. (2021). Convolutional Neural Networks based classification of breast ultrasonography images by hybrid method with respect to benign, malignant, and normal using mRMR. *Computers in Biology and Medicine*, *133*. https://doi.org/10.1016/j.compbiomed.2021.104407

Gu, Y., Xu, W., Lin, B., An, X., Tian, J., Ran, H., Ren, W., Chang, C., Yuan, J., Kang, C., Deng, Y., Wang, H., Luo, B., Guo, S., Zhou, Q., Xue, E., Zhan, W., Zhou, Q., Li, J., … Jiang, Y. (2022). Deep learning based on ultrasound images assists breast lesion diagnosis in China: a multicenter diagnostic study. *Insights into Imaging*, *13*(1). https://doi.org/10.1186/s13244-022-01259-8

Meraj, T., Alosaimi, W., Alouffi, B., Rauf, H. T., Kumar, S. A., Damaševičius, R., & Alyami, H. (2021). A quantization assisted U-Net study with ICA and deep features fusion for breast cancer identification using ultrasonic data. *PeerJ Computer Science*, *7*. https://doi.org/10.7717/PEERJ-CS.805

Sahu, A., Das, P. K., & Meher, S. (2023). High accuracy hybrid CNN classifiers for breast cancer detection using mammogram and ultrasound datasets. *Biomedical Signal Processing and Control*, *80*. https://doi.org/10.1016/j.bspc.2022.104292

Singh, P., Mukundan, R., & De Ryke, R. (2020). Feature Enhancement in Medical Ultrasound Videos Using Contrast-Limited Adaptive Histogram Equalization. *Journal of Digital Imaging*, *33*(1), 273–285. https://doi.org/10.1007/s10278-019-00211-5

Thomas, C., Byra, M., Marti, R., Yap, M. H., & Zwiggelaar, R. (2023). BUS-Set_A benchmark for quantitative evaluation of breast ultrasound segmetnation networks with public datasets. *Medical Physics*, *50*, 3223–3243.

Wang, Y., & Yao, Y. (2022). Breast lesion detection using an anchor-free network from ultrasound images with segmentation-based enhancement. *Scientific Reports*, *12*(1). https://doi.org/10.1038/s41598-022-18747-y

Zhang, M., Bao, H., Zhang, X., Huang, Z., Zhao, Z., Li, C., Zhou, M., Wu, J., Wang, L., & Wang, L. (2018). *Preplanned Studies Breast Cancer Screening Coverage-China*.

11.2 Subject Matter Expert Interview: Questionnaire & Responses

Physician responses (translated by sponsor) are shown in italics.

16Jun2023

1. Can you briefly describe your workflow, from starting to take the ultrasound, to deciding on a diagnosis?

In general, the operating doctor will first inquire about the patient's medical history and observe if there are any incisions or open wounds on the patient's body. The assistant will check the patient's medical history, CT scans, X-rays, and other examination results in the medical record system. When starting the examination, the operating doctor uses a probe to scan all sections of the organ to be examined.

2. Is the physician who operates the ultrasound machine the same person who reviews the images and makes a diagnosis?

It's the same doctor. The operating doctor is responsible for performing the operation, examining the images, and making a diagnosis.

3. What are your biggest challenges that limit your confidence when making a diagnosis based on ultrasound imaging?

The final pathological result was different from the diagnosis made by the ultrasound doctor; it is difficult to diagnose a disease when it has multiple manifestations; adjacent organs can cause confusion, such as mistaking the intestine near the gallbladder for the gallbladder.

4. Is the BI-RADS ranking system used to classify diagnoses in your hospital?

Yes, we are using the Bi-RADS ranking system.

5. Are there other patient factors you consider when making a diagnosis (such as age, family history of cancer, physical condition) in addition to the ultrasound images?

We will take age into consideration. If a lump is discovered for the first time in a person above the age of 45, we may upgrade the BIRADS rating (for example, a patient who is initially rated as grade 3 based on image features may be ultimately rated as 4a grade). Family history is not given special attention in the ultrasound department. The most important aspect is still observing the morphological features of the image.

6. Are there other attributes of the lesion that you consider when making a diagnosis (such as tumor size or location) in addition to the ultrasound images?

For the breasts, size and location do not have a special influence on the doctor's judgment. Particularly large masses may also be benign.

7. For a given patient, how many different views or images (for example, longitudinal and transverse planes) do you typically examine before making a diagnosis?

There are multiple views during breast examination, especially for sagging breasts. If the ultrasound probe slides too quickly, it is easy to miss something. Generally speaking, I will perform transverse, longitudinal, and oblique cuts, and if a certain area is suspected to be suspicious, I will scan it multiple times.

8. What do you think is the false-positive diagnosis rate (diagnosis of malignant, which is proven to be benign after biopsy) for human experts using only ultrasound imaging?

I cannot answer this question. It may require big data statistics to draw conclusions. In our hospital, we recommend patients with a breast imaging reporting and data system category of 4b or above to undergo biopsy and patients with a category of 4a to have a follow-up check-up after three months.

9. If a high-accuracy image analysis tool was available to support you in making a diagnosis, what kind of output would be helpful? (Select all that apply, and make any comments you like)
    a. Enhanced ultrasound image
    b. Annotated ultrasound image showing the region where the tool has detected a tumor
    c. Clear recommended binary diagnosis: benign/malignant
    d. Percent probability or confidence of the recommended diagnosis
    e. Recommended BI-RADS ranking
    f. Other (please specify):

I think a,b,d,e would be helpful.
c: If provided for doctors as reference, it's fine. But it's not recommended to tell patients directly as it may cause disputes.
d: I think defining the confidence interval will be very challenging.

10. Is there anything else you think the team should know, when developing a machine learning tool to support your work?

I think it's necessary to have a good understanding of the relevant medical knowledge related to breasts, such as their structure, tissue organization, and lymphatic metastasis.