Final Report - Team 8: Perimeter AI

Anne Jing, Jade Clement, Daniel Deza, Dhruv Sirohi, Jim Yang, Tony Chung

I - Introduction

A) Problem Statement

Breast cancer is the most common cancer worldwide and the leading cause of cancer-related deaths among women [1]. A lumpectomy is a procedure that removes a breast tumor but may leave behind residual tumor cells [2]. Perimeter Medical Imaging (PMI) has implemented AI to flag tumors on OCT scans[1] of breast tissue. This accelerates the process of finding suspicious tissue from several days of labour to mere minutes. However, Perimeter AI lacks a range of cancer-positive labelled data to train their model and the labelling process of scans containing cancerous cells is time-consuming as it requires the judgement of two experts. Indeed, despite collecting nearly 15 million patches, only 0.36% or 53 thousand of these patches are labelled. Furthermore, among these labelled patches, the negative patches dwarf the positive patches with nearly one positive patch for every four negative ones. As such, Perimeter AI is seeking machine learning methods to acquire additional positive labelled data.

While the issue is clear, there is a lack of positive labelled data hindering the performance of Perimeter AI, the reasons behind it need analysis. Below we can see the data collection pipeline for Perimeter AI.

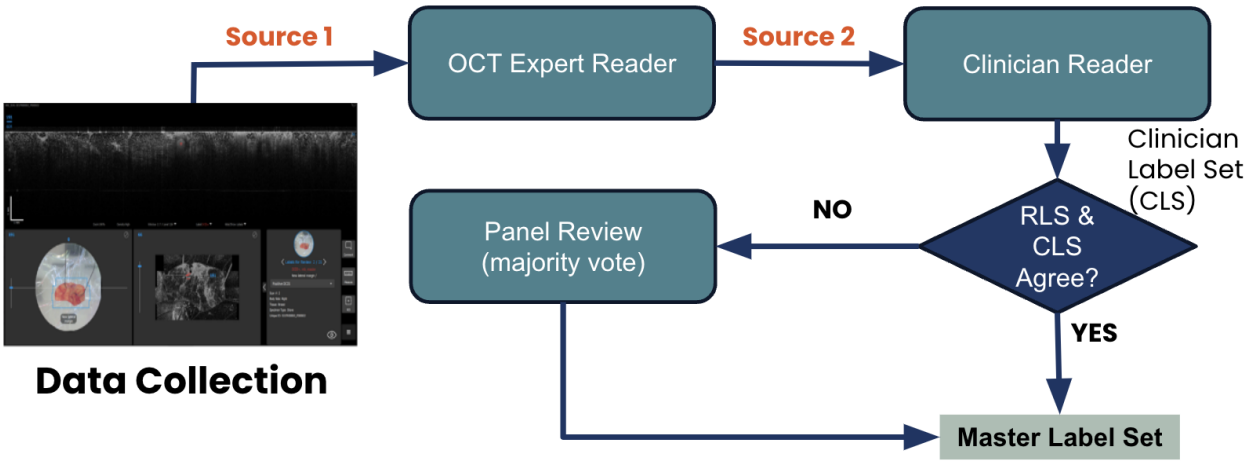


Figure 1: Data collection pipeline

OCT scans are taken and then labelled by two expert readers, if they can identify the tumor on the scan then it can be labelled as positive. Consulting with the clients, we can isolate two sources of the lack of positive data.

[1] A high-resolution imaging technique that assesses surgical breast tumor margins

(1) The first, more intuitive, source is that there are just more cancer free people than cancerous people.
(2) The second source is that around 50% of the time, despite the patch being suspicious, the expert readers are unable to pinpoint the tumor on the patch and therefore cannot label it as positive.

B) Requirements

To address these different sources we identified two set approaches: generating new positive data from the existing data and assisting the expert readers with locating the tumors on suspicious data. The requirements for these two approaches are outlined below and will be further explored in the Methodology section.

Approach	Objectives	Constraints/Requirements	Metrics
1. *Data augmentation techniques*	Increase the amount and variability of positively labelled data within a dataset	Should not be computationally demanding	The less CPU usage the better
		Must increase the area of the Precision-Recall curve	Improvement of 5% to 10% from the current 0.81.
2. *Automated assistance of the expert labelling*	Highlighting suspicious areas on scans to increase the speed and performance of the labeller	Should not be computationally demanding	The less CPU usage the better
		Should single out subsections of the scan where the cancer is suspected to be	Binary: [Y] can single out [N] cannot single out
		Accuracy should outperform baseline[2] accuracy (50%)	>50% accuracy

Table 1: Proposed approaches and requirements

II - Data

Perimeter AI provided us with nearly 50 thousand black & white labelled patches comprised of 50% positive data and 50% negative data. This data was split into eight subfolders according to what the content of the patch was (cyst, duct, bloodvessel), three of which corresponded to positive data folders split based on how easily the expert readers were able to spot the tumors. We then reorganised the data into two folders; positive and negative. An example of a positive and negative patch are shown below.

[2] Manual inspection and labelling by expert

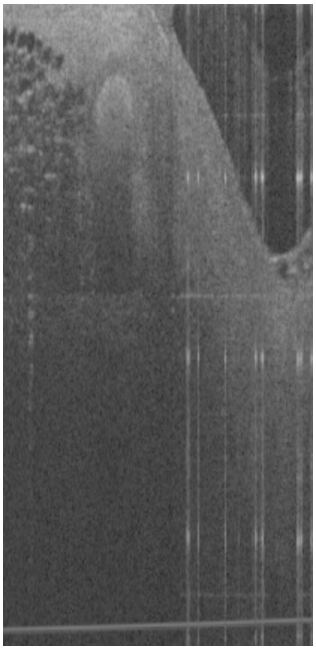


Figure 2.a Positive Data Sample

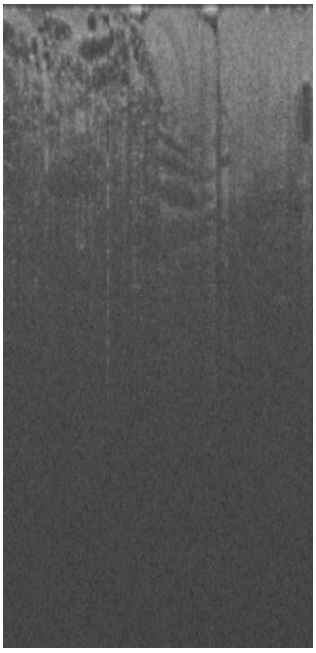


Figure 2.b Negative Data Sample

This data was split into train, test, and validation sets with 75%, 15%, and 10% of the data respectively. Furthermore, Perimeter AI provided 36 patches with coordinate labels on each image identifying the different parts of the scan, such as the coordinates of the tumor, the ducts, lipids etc. This data was used for the labelling assistance approach (*Approach 2*)

III - Methods

Perimeter AI has highlighted the importance of exploring a breadth of methods as even those that do not come to fruition aid them in making more informed decision on where to invest their time and resources next. Thus, our methodologies have been split into two approaches to tackle both insufficient sources of positive data. The first set of approaches focuses on creating new positive data through transformation of existing positive data or by generating new images entirely. The second helps the expert labeler by pointing out where the tumor cell is likely to be on the scan, increasing their accuracy and speed.

Approach 1 - Data augmentation techniques

A) Elastic Transform

This data augmentation technique involves deforming an image by stretching and compressing it in random directions. These kinds of deformations are similar to what happens to the tissues

when they are prepped for analysis. Thus, if the model learns from this augmented data, it can become more robust to these patterns. Elastic transformation is achieved by adding small displacements to each pixel then using a Gaussian filter to smooth it out. The strength of the transformation is determined by two hyperparameters: alpha defines the intensity of the distortion and sigma controls the amount of smoothness between the pixels.

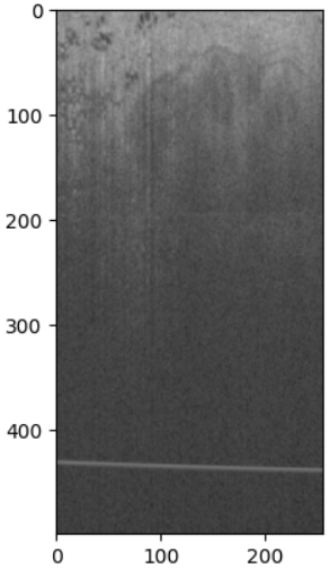


Figure 3.a Original Data Sample

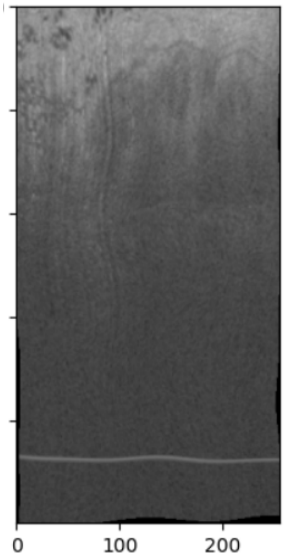


Figure 3.b
Transformed Data Sample
alpha=1000, sigma=40

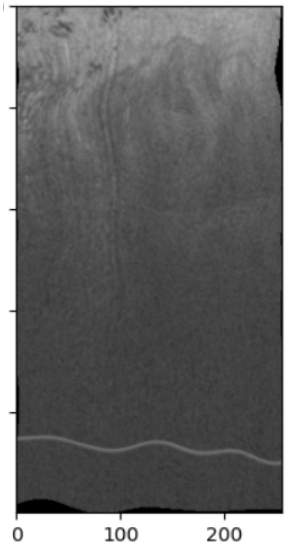


Figure 3.c
Transformed Data Sample
alpha=600, sigma=20

Figure 3 demonstrates two different transformations of an example original OCT scan. It is crucial that the changes are noticeable enough to count as a new data point so that the model can learn something new. However, transforming the image too much would be detrimental because it would then lose the features that define the cancer and the model would be training on incorrect data. Figure X.b represents a well transformed data point that achieves an optimal balance between these two requirements. From visual inspection it is clear that Figure X.c is poorly transformed, many hyperparameters had to be tested to find optimal ones to use for training.

B) Filtering

Filtering involves applying various image processing filters to the original images to create modified versions, thereby simulating a range of possible imaging conditions and scenarios. Two filtering methods were used for data augmentation: Gaussian Filtering and Median Filtering.

B.1) Gaussian Filtering

Gaussian filtering in image processing is used to smooth images and reduce noise. This technique involves convolving an image with a Gaussian kernel. The Gaussian kernel is a matrix that gives more weight to pixels closer to the centre of the window, following the bell-shaped curve of the Gaussian function. The size of the kernel is a measure of its width and height and can be hypertuned. A larger kernel covers more pixels in its operation, thereby averaging over a larger area which results in a more blurred image. In our experiments of data augmentation, we applied a Gaussian filter to cancer-positive images using a kernel size of 5x5 which was implemented using the GaussianBlur function in OpenCV. This size was determined visually as a kernel size too small would not produce an image unique enough and a kernel size too large could result in the loss of important details and features in the images. The computational cost of applying a Gaussian filter is relatively inexpensive in computational resources, especially in comparison to other, more complex image processing operations; this efficiency makes it a popular choice for data augmentation.

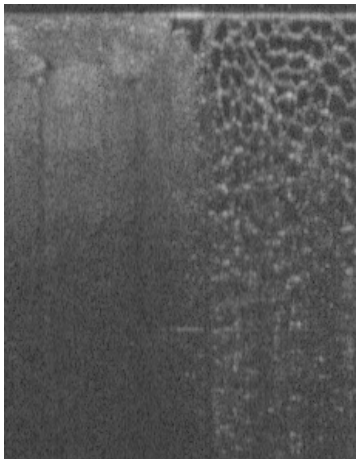


Figure 4.a
Original Data Sample

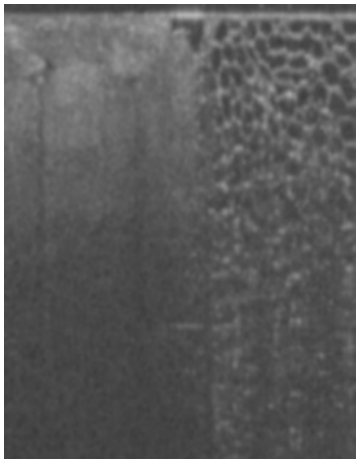


Figure 4.b
Gaussian Blurred Data Sample, kernel size = 5x5

B.2) Median Filtering

Unlike the Gaussian blur, which replaces each pixel value with a weighted average of surrounding pixel values, the median filter replaces each pixel's value with the median value of its neighbouring pixels. One of the primary advantages of the median filter is its ability to remove noise without significantly blurring the edges of objects in an image. Since the median is less sensitive to outliers than the mean, when the filter encounters a pixel significantly different from its neighbours, it is replaced with the median value, which is more representative of the area than the mean. The aperture size determines the number of neighbouring pixels considered when calculating the median value. In our experiment, we used an aperture size of 5x5 which was implemented using the medianBlur function in OpenCV.

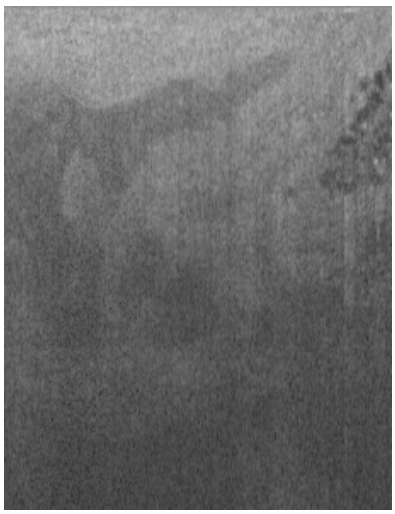


Figure 5.a
Original Data Sample

Figure 5.b
Median Blurred Data Sample,
Aperture size = 5x5

C) Masking

Image masking, more specifically, unsharp masking is a method to highlight edges and interfaces to make the image crisper and more defined. Such details are achieved by taking the difference between the original image and its blurred version, mainly via Gaussian blur, where then the details are scaled and added back to the original image.

We utilised unsharp masking via unsharp_mask function in Scikit-image library, in which the resulting images were fine-tuned and fed into our model for evaluation.

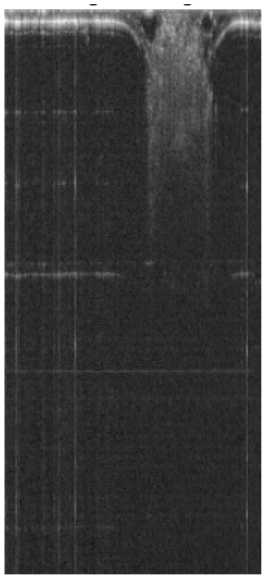


Figure 6.a
Original Data Sample

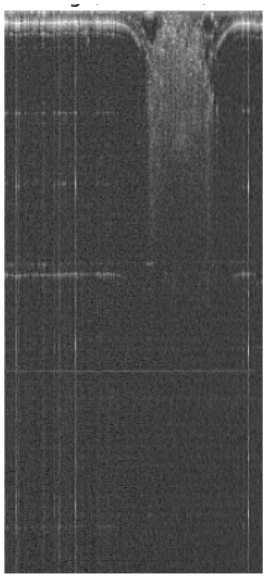


Figure 6.b
Masked Data Sample,
Radius = 1, amount = 1

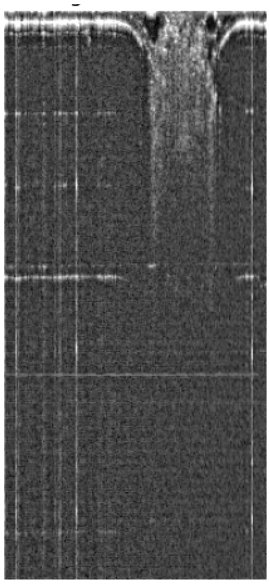


Figure 6.c
Masked Data Sample,
Radius = 5, amount = 2

We deployed various parameters during image masking, specifically radius and amount parameters. Radius parameter refers to the sigma or variance parameter of the gaussian filter, which dictates how much the image is blurred before being subtracted from the original image to make it sharper. The amount parameter scales the amount of blurring in which is subtracted from the original image.

The objective of these data augmentation techniques was to create more positively-labelled data for our model to train on. While the three techniques apply different image processing algorithms, we were able to achieve enough positive data to enhance training and validation accuracy of our model in classifying scans with cancer tumors.

D) CGAN

This method explores augmenting the collection of labelled positive image data by generating entirely new data from the existing image data. In order to do this, we looked at Cycle GANs which involved building a discriminator to differentiate between positive and negative samples, and a generator to create images conditioned on a given label and then training these two models in a loop. Ultimately we found this model to be very computationally expensive while also not showing promising results despite running for many epochs. Considering the time investment and our goal of exploring a breadth of methods we didn't explore it further for our project. However, it is still a viable path that Perimeter AI could explore.

Approach 2 - Automated assistance of the expert labelling

A) Object Detection

Object detection models were explored to aid during the expert labelling portion of the data process. This could be done by suggesting regions of interest for each image that experts should investigate in depth. If these suggestions are done accurately, it would be possible to dramatically speed up the expert labelling task, and also potentially reduce the number of voided cancer-positive data points. The key metric of importance here is recall: high recall is particularly crucial in medical image analysis because the cost of missing a malignant tumor is much higher than the cost of over-flagging potential areas of concern. Such models were chosen because they have been shown to perform well on similar tasks in literature, even when there is a limited number of labelled images.

The main technique explored for this problem was using a YOLOv5 object detection model trained on a subset of 36 images from the main dataset for which the Perimeter team provided the coordinate labels needed for object detection. Due to the precise nature of this problem, it is not possible for us to generate such labels for unlabelled images in the dataset on our own - it is necessary to receive verified coordinate labels from Perimeter. In this case, the labels are

coordinates denoting the center of tumors in each positive scan. This limited dataset was expanded through simple x and y-flip augmentations to a total of 108 images.

Over training the YOLOv5 model across 15 epochs, classification loss on the training set drastically decreases, indicating that the model possesses the necessary complexity and capacity to handle the nuances of the dataset it has been presented with. Moreover, model recall generally increases throughout training, reaching a max of 0.8. This is important, as it suggests that the model is not missing many areas of importance in its suggestions. Adding significantly more labelled images to the dataset used for this training will be necessary if we wish to improve the model's performance to a level sufficient for expert assistance.

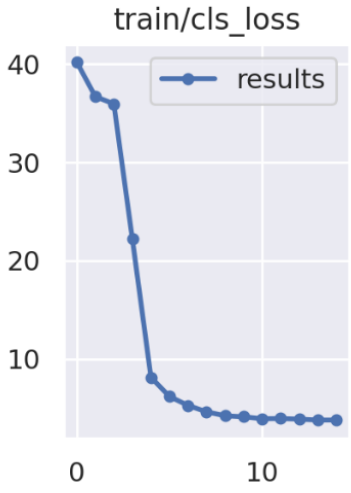


Figure 7. Classification Loss per Epoch

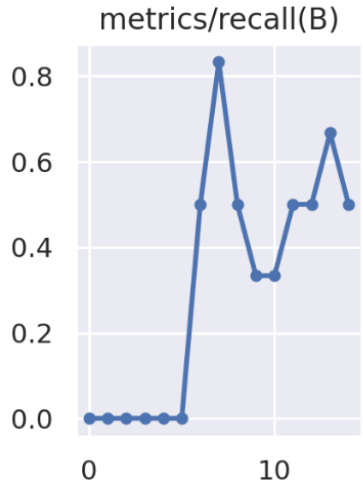


Figure 8. Model Recall on Dataset across Epoch

IV Results

A) Evaluation Method

An effective augmentation method ensures that the derived images retain notable characteristics of cancerous imagery while also diversifying the features embedded within the images. This diversity infuses the dataset with more information, thus facilitating a more robust learning process for the machine learning model. Our client, Perimeter AI, has their proprietary machine learning model for the classification of positive and negative images that is very resource intensive. For our purposes, they have recommended we opt for ResNet18 [4] as a proxy for the company's classification model. In line with the client's suggestion, our team has effectively adopted ResNet18 in the capacity of a binary classifier. This allowed us to compare the test accuracies of the same model on each dataset.

B) Training Process

For testing, we first fine tuned the base ResNet18 model on ~8200 original images. We then trained the model on 4 additional datasets:

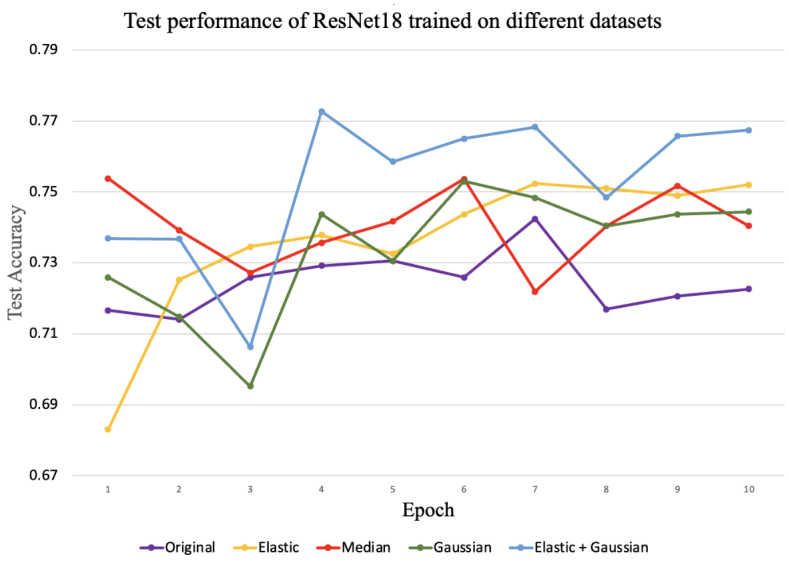
- Dataset 2: ~8200 original images + ~4000 elastic transformed images
- Dataset 3: ~8200 original images + ~4000 median filtered images
- Dataset 4: ~8200 original images + ~4000 gaussian filtered images
- Dataset 5: ~8200 original images + ~4000 elastic transformed as well as gaussian filtered images

Note that we also tested other combinations of these methods but only kept the better performing ones. Each augmented dataset represents a combination of the original images and those images generated through a specific augmentation method integrated into a singular dataset. The original image dataset encompasses both negative and positive data, whereas the augmented images are exclusively positive, aligning with the client's specific need for more positive labelled images. While it may seem unfair to train the model on larger datasets than the base dataset, considering we are testing data augmentation techniques this reflects how Perimeter AI will use the data in practice.

C) Testing

The graph below summarises the testing accuracies of the models trained on each dataset per epoch. The testing was done on 1700 original images not in the training or validation set. We can see that the original data set reaches a test accuracy of 72% while our best performing model, elastic transform with gaussian filtering, reaches a test accuracy of 77%.

Figure 8. Test Accuracy During Each Epoch



V Discussion & Implementation

While the testing done is only on a proxy of Perimeter AI's model as well as not being able to empirically measure the performance increase that the object detection approach would offer, the methods developed in the report do address both the sources of inefficiencies outlined by Perimeter AI. This is significant as it prevents future bottlenecks where despite major improvements at one stage, the slow pace at the other stage slows the whole process down. Beyond just the performance increase discussed previously, we also met the goals of the stakeholders. We explored a breadth of solutions, even though not all panned out they do serve as a guide as to where to focus their resources in the future. Furthermore, the successful models we did put forward remain relatively computationally inexpensive.

The lack of positive data is an issue that plagues much of the medical field and since our models are generalizable to different types of image data the work we did could be useful to the medical industry in general. Finally, there are some limitations to the solution put forward. The most evident is that Perimeter AI is still reliant on collecting new data as the models are transformative and not generative. The second is that the object detection model is trained on highly annotated data that is time consuming to create manually meaning while it lessens the burden on the current data collection and labelling pipeline it also adds a new time consuming step. Yet, there are implementation methods Perimeter AI could explore to address the latter. To create more labelled data to further train the objection detection they model they could set up a loop where the output of the model is then fed back into the model as new data. At first this would require more intense checking and correction of the outputs but as the model refines itself less and less human intervention would be required.

VI Conclusions and Future Directions

To summarise, the solution we propose to aid Perimeter AI with their lack of positive data is two fold. The first is an elastic transformer with a gaussian filter to create new positive data from the existing data. The second is an object detection model which helps their expert labels with identifying positive data. Reflecting upon the initial requirements, both models are rather computationally cheap while showing a promising performance. In the future, we recommend that Perimeter AI uses the objection detection model alongside human correction to create more highly labelled data to further train the model. Finally, considering our goal of exploring a breadth of solutions we did not have sufficient resources to properly investigate CGANs but we recommend Perimeter AI to take another look at it as well as other generative models as that would alleviate their reliance on collecting data.

Attribution Table

While every member had their hand in each part of the final report and project in general, the table below is indicative of the team leads / experts for each part.

Team Member	Contribution
Anne Jing	● Background research into cancer ● Determined requirements ● Researched various methods of filtering. Specifically: ○ Gaussian Filtering ○ Median Filtering ○ Gabor Filtering ○ Sobel Filtering ○ High-boost Filtering ○ Laplacian Filtering ● Used decision matrices to determine which filtering methods to pursue due to limited time and resources ● Implemented Gaussian Filtering and Median Filtering for all cancer-positive images ● Tuned hyperparameters of the Gaussian Filtering and Median Filtering to achieve optimal results
Jade Clement	● Elastic Transform, CGAN, Data cleaning
Daniel Deza	● CGAN ● Presentation slides ● communication with Perimeter AI ○ Email contact with Perimeter AI ○ Presented our progress during meetings
Dhruv Sirohi	● Object Detection, Labelled Data collection / handling
Jim Yang	● ResNet18 testing and tuning, Data pipeline
Tony Chung	● Masking, Filtering

Citations

[1] “Breast cancer,” World Health Organization, https://www.who.int/news-room/fact-sheets/detail/breast-cancer#:~:text=Scope%20of%20the%20problem,the%20world’s%20most%20prevalent%20cancer. (accessed Nov. 3, 2023).

[2] “Breast-conserving surgery (lumpectomy): Treating breast cancer,” Treating Breast Cancer | American Cancer Society, https://www.cancer.org/cancer/types/breast-cancer/treatment/surgery-for-breast-cancer/breast-conserving-surgery-lumpectomy.html#:~:text=Have%20a%20tumor%20smaller%20than,risking%20harm%20to%20the%20fetus) (accessed Nov. 3, 2023).

[3] Sangeeta Rani, Bhupesh Kumar Singh, Deepika Koundal, Vijay Anant Athavale, Localization of stroke lesion in MRI images using object detection techniques: A comprehensive review, Neuroscience Informatics, Volume 2, Issue 3, 2022, 100070, ISSN 2772-5286

[4] https://pytorch.org/vision/main/models/generated/torchvision.models.resnet18.html