Augmenting Deep Learning to Improve Breast Cancer Detection through Mammography

Ambri Ma[1] and Arnav Kumar[2]

[1]Department of Operations Research and Financial Engineering, Princeton University
[2]Department of Computer Science, Princeton University

Advised by William Yang
COS429 - Computer Vision
Taught by Prof. Olga Russakovsky
Department of Computer Science, Princeton University

Project Codebase: https://github.com/ambrim/breast_cancer_detection

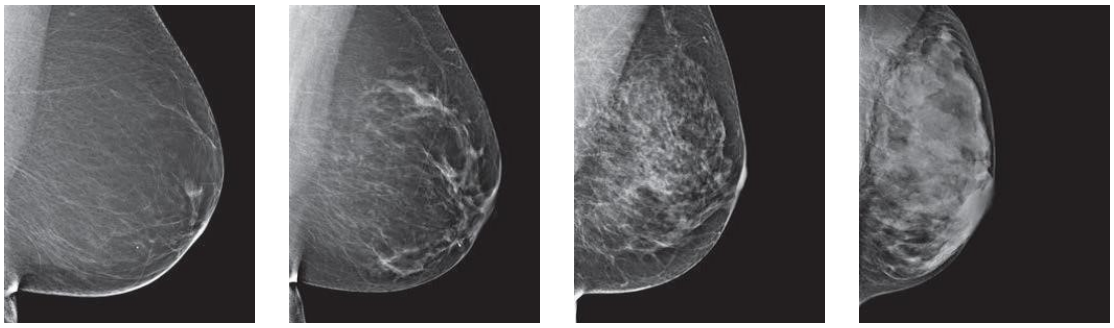
Contents

1 Introduction: Background and Motivation

Breast cancer is the most commonly occurring cancer globally, with over 2.26 million cases and almost 685,000 mortalities worldwide in 2020 [3]. Around 1 in 8 women in the United States will develop invasive breast cancer in their lifetime[2]. In 2022, nearly 290,000 invasive cases and 52,000 non-invasive cases were diagnosed in the US alone [2]. With such high prevalence, early detection of breast cancer is crucial for effective treatment and improved patient outcomes. Only 10% of breast cancers are linked to genetic mutations inherited from a parent; the other 85% of cases occur in women with genetic mutations from aging and no family history of breast cancer [2]. As a result, the risk of breast cancer increases greatly as one gets older.

Medical imaging, such as mammography, is the primary method for breast cancer screening and diagnosis. As the first step, a screening mammogram is used to detect abnormal growths or changes in breast tissue of people who are at risk but may not exhibit breast cancer symptoms [9]. X-ray pictures of each breast are taken, typically from two angles, although more angles may be incorporated for diagnosis mammograms. These pictures may find calcifications, masses, asymmetries, and other distortions used in diagnosis and protocols for further testing, such as biopsy [9]. Typical diagnostic methods include annual screenings for women over the age of 45 via mammograms (low-dose X-rays) and clinical exams. If any sort of irregularity is found during a screening, an ultrasound, MRI (Magnetic Resonance Imaging), or biopsy may be used to further evaluate the tissue condition and inform treatment plans [9].

However, human interpretation of medical images is highly subjective, tedious, prone to human error, and presents many challenges, even for experienced radiologists. A 2016 study at MGH found that the sensitivity, accuracy among true positives, of mammography detection is around 87% and the specificity, accuracy among true negatives, is around 89% across 359 radiologists at 95 facilities [6]. False positive results can lead to extra testing and cause anxiety, whereas false negative results will delay diagnosis and treatment [9]. Moreover, false negative results are more likely for younger patients with denser breast tissue and cancer that is fast-growing [9]. Refer to Figure 1 for example X-rays of breasts with varying densities.



(a) Least Dense: Breasts are almost all fatty tissue.

(b) Less Dense: Scattered areas of dense glandular and fibrous tissue, or heterogeneously dense tissue, are seen as white areas on the mammogram.

(c) More Dense: More of the breast is heterogeneously dense. Masses in or around the dense tissue are hard to detect since they also appear as white masses.

(d) Most Dense: Breasts are extremely dense.

Figure 1: Four categories of breast density as determined by mammography. Image credit to the American Cancer Society [1].

There are several different types of breast tumors, often differentiated by their behaviors or repercussions on the body. A benign tumor is an abnormal but noncancerous collection of cells [4]. It can form anywhere in the body when cells multiply more or die at slower rates than expected. Benign tumors grow slowly, do not spread to other parts of the body, and do not require treatment. A malignant tumor grows at a much faster rate than other functional cells in order to invade nearby tissue, enter the bloodstream, and proliferate in other body parts. In the malignant case, breast cancer stages T1-T3 are determined by the tumor size, with stage T4 signifying when the tumor has grown into other parts of the body [9]. One of the main

physical differences used to aid detection of benign vs. malignant tumors is the shape of the irregularity; benign tumors typically exhibit a smooth, regular, and circular border, whereas malignant tumors are more irregular and indistinct [4].

Early detection of breast cancer greatly increases the likelihood of successful treatment and survival, as cancer metastasizes and becomes more challenging to manage as it progresses. Although mammograms are widely used and trusted in detecting breast cancers, it is estimated that a review by one radiologist can miss between 16 to 31% of cancers that are detectable via mammogram [6]. Although the likelihood of diagnosis increases with a second physician, so does the false positive rate. Computer-assisted detection and diagnosis (CAD) software has been developed since the 1990s to enhance the predictive accuracy of screening mammography for radiologists [11]. However, early commercial CAD systems did not yield substantial improvements in performance, and progress remained stagnant for over a decade since their introduction. With the notable success of deep learning in visual object recognition, detection, and various other domains, there is growing interest in developing deep learning tools to support radiologists and enhance the accuracy of screening mammography [11]. Recent studies have demonstrated that a deep learning-based CAD system performed comparably to radiologists in standalone mode and improved their performance in support mode [11, 10].

2 Related Work

The rapid progress of machine learning and especially deep learning has contributed significantly to the medical imaging community's growing interest in leveraging these techniques to enhance the cancer screening process. Recent intelligent classifiers based on conventional machine learning algorithms have the potential to improve the accuracy and speed of breast cancer detection. With convolutional neural networks, researchers at Mount Sinai medical school have achieved results of 87% sensitivity and 96% specificity with an "end-to-end" training approach, surpassing doctors in avoiding false negative diagnoses [11]. In their approach, lesion annotations were required only in the initial training stage, and subsequent stages required only image-level labels, eliminating the reliance on rarely available lesion annotations.

However, these statistics overlook the specific issue of age, which may significantly impact accuracy of breast cancer screenings [8]. In particular, younger people (ages 25-49) exhibit starkly higher breast density than older patients (49+), introducing motivation to improve screening accuracy for younger populations. Moreover, people are less inclined to get screened for breast cancer at younger ages, as people with ages 30-49 have the lowest participation rates in screening mammography out of a population of ages 30-75 [8].

The medical applications of age estimation have garnered significant attention, and although numerous studies have explored the use of biomedical images to estimate human age, none had bridged the gap between age and mammography-based breast cancer detection. The study which originally published the mini-DDSM dataset addressed this gap by developing an AI-based model for estimating age from mammogram images [7]. To do so, researchers extracted deep learning ResNet features from the collected dataset and used a Random Forests regressor model to automatically estimate age. The authors evaluated model performance with mean absolute error, with an average error value of approximately 8 years obtained from 10 tests on randomly selected samples.

For the purposes of computer-aided disease diagnosis classifying malignant and benign tumors from mammography, we aimed to apply the full-image classification task proposed in [11] to the dataset published in [7] in order to potentially augment data for younger populations. We modify the 5-way classification problem in [11] to the 3-way classification problem due to the labeling constraints of the mini-DDSM dataset. Given the success demonstrated by previous studies in end-to-end full-image classification, our aim is to extend the investigation into potential inaccuracies in breast cancer classification for younger populations.

2.1 Dataset: Mini-DDSM

The mini-DDSM dataset is a lighter version of the formerly popular DDSM (Digital Database for Screening Mammography, now obsolete, lossless-JPEG format) and the CBIS-DDSM (Curated Breast Imaging Subset of DDSM, 163.6 GB, DICOM format), which are databases of scanned film mammography studies. Among other extensively researched datasets such as the Mammographic Image Analysis Society (MIAS) database, only the DDSM contains the age attribute for each image. In summary, the mini-DDSM database

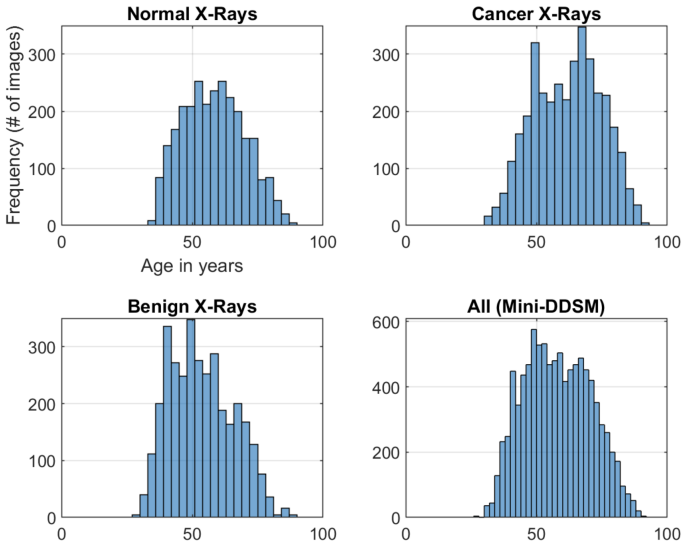


Figure 2: Age distribution in the Mini-DDSM dataset

contains examples from three conditions: 2,728 images in the Normal (no tumor) category, 3,596 images in the Benign (tumor) category, and 3,360 images in the Malignant (tumor) category, for a total of 9,684 images. The mini-DDSM includes both craniocaudal (CC) and mediolateral oblique (MLO) views from the Left and Right sides for most exams. Each perspective was treated as a separate image in this study. Refer the age distributions for each category in Figure 2.

Medical imaging data for patients aged below 40 and those above 50 exhibit stark differences, posing a challenge for accurate breast cancer classification in younger populations. In this study, we aim to investigate potential inaccuracies associated with age-related differences in mammograms, hypothesizing that such differences will lead to lower classification accuracies among younger age groups. We will analyze the results of our baseline model and use them to inform the design of an augmented model to adjust for disproportionate data. Our goal is to create a more unbiased classifier that can better account for differences between age groups.

2.2 Computational Environment

All experiments in this study were carried out on a Macbook Pro workstation equipped with a 2 GHz Quad-Core Intel i5 CPU and on a Macbook Air workstation equipped with a Apple M1 chip.

3 Model Architecture, Parameters, and Baseline Performance

In this section, we evaluate three different model architectures to create our baseline: a SimpleCNN, a pre-trained VGG-16 CNN, and a pre-trained ResNet-50 CNN. The objective is to obtain a reliable model to investigate potential age bias in breast cancer classification. The baseline model selected in this section will serve as a reference for the remainder of the study.

We split the mini-DDSM dataset randomly into training and test sets in an 80:20 ratio, ensuring that the different age group proportions were maintained using stratified sampling. Cross-validation was employed in a 90:10 ratio within the training set.

We trained all models using identical hyperparameters, which were selected based on validation performance. Adam was selected as the optimizer, employing a learning rate of 10^{-5}. However, due to time and resource limitations, we were restricted with the scale of total epochs, batch size, and steps per epoch. Within reasonable boundaries, we set the batch size to 20, the number of total epochs to 50, and the number of steps per epoch to 200. No data augmentation was added to the baseline models.

3.1 SimpleCNN

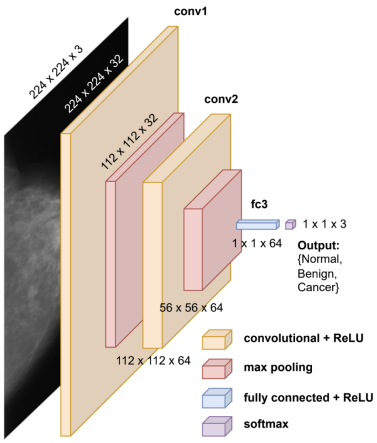


Figure 3: Network Architecture implemented with a simple CNN

The first model architecture we explored was a simple CNN, depicted in Figure 3. Its architecture consisted of two convolutional layers with 32 and 64 filters, followed by max pooling layers with a 2x2 pool size. The output from the convolutional layers was flattened and inputted into a fully connected layer with 64 units and ReLU activation. Finally, an output layer comprising of softmax activation was incorporated to output three classes corresponding to the three categories of scans: "Normal," "Benign" (referring to tumor or possibly calcification), or "Malignant" (referring to tumor or calcification). We selected this architecture based on its positive results and relatively efficient training time compared to other simpler and more complex architectures after trial-and-error. Test set evaluation yielded the following results, shown in Table 1.

Class	Accuracy	Precision	Recall	F1-Score
Normal	0.70	0.87	0.70	0.78
Benign	0.55	0.60	0.55	0.57
Cancer	0.71	0.56	0.71	0.65
Overall	0.65	0.67	0.65	0.65

Table 1: SimpleCNN Results

The SimpleCNN achieved an overall accuracy of 64.7%. Further details of the model's performance are displayed in Figure 7. From the confusion matrix, we discern that the model exhibited adequate accuracy in discriminating images based on the presence or absence of a tumor. The model was able to identify 70% of scans without a tumor correctly, and it detected the presence of a tumor or calcification in 94% of benign scans and 97% of malignant scans. However, the model struggled to differentiate between benign and cancerous tumors, achieving an accuracy of only 55% on benign tumors despite a respectable 71% accuracy on malignant tumors. Although the overall results are sufficient for the sake of our initial hypothesis, we also explored the potential benefits of transfer learning using pretrained ResNet-50 and VGG-16 models, detailed in the following sections.

3.2 ResNet-50

Our second model architecture, as illustrated in Figure 4, utilized a pretrained ResNet-50 model as the feature extractor. To fine-tune the network, we added two convolutional layers with ReLU activation of 512 and 256 filters, respectively. Between these convolutional layers, we applied max pooling with a 2x2 pool size. The resulting output was then flattened and fed into four fully connected layers of size 256 with ReLU activation. To prevent overfitting to the training data, we added a dropout layer. Finally, a dense layer with

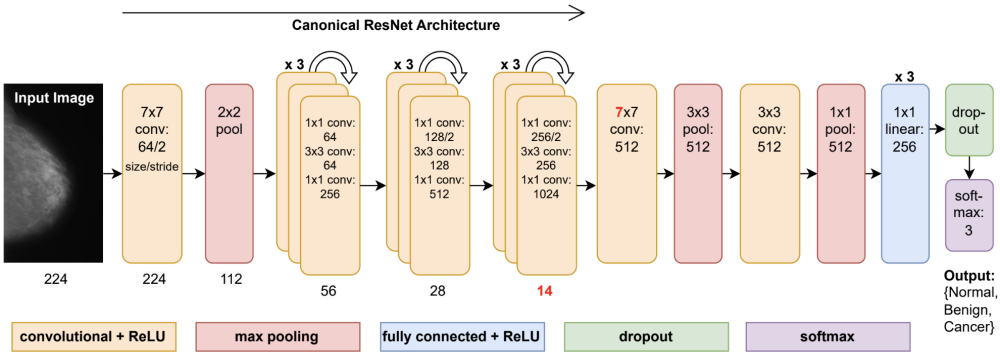


Figure 4: Network Architecture implemented with ResNet-50

softmax activation of size 3 was employed to generate the output. The results after 50 epochs are the shown in Table 2:

Class	Accuracy	Precision	Recall	F1-Score
Normal	0.56	0.96	0.56	0.71
Benign	0.08	0.52	0.08	0.14
Cancer	0.95	0.42	0.95	0.58
Overall	0.52	0.61	0.52	0.45

Table 2: ResNet-50 Results

The ResNet-50-based network demonstrated overall accuracy of 51.5% after 50 epochs. This accuracy is notably 13.2 percentage points lower than the accuracy achieved by SimpleCNN. As depicted in Figure 7, the ResNet-50 model detected tumors at a rate of 99% for both benign and malignant scans, but exhibited a much lower rate for normal scans at just 56%. Moreover, the model displayed weak performance in classifying tumors as malignant or benign, misclassifying 91% of benign tumors as malignant. It is surprising that the ResNet-50 pretrained model revealed poor results compared to our SimpleCNN setup. This discrepancy may be explained by the short training duration, as 50 epochs with 200 steps each was perhaps not sufficient for the model to converge, as supported by Figure 6.

3.3 VGG-16

Our third network architecture was implemented using a pretrained VGG-16 model as the feature extractor, followed by an identical fine-tuning process as before: via the addition of two convolutional layers with ReLU activation using 512 and 256 filters, respectively, and max pooling outputs with a 2x2 pool size. The resulting output was then flattened and received by four fully connected layers, each with size 256 and ReLU activation. To mitigate overfitting, the output from these layers was then passed through a dropout layer, followed by a final dense layer employing softmax activation of size 3 for generating the output. The results obtained after training for 50 epochs are found in Table 3.

Class	Accuracy	Precision	Recall	F1-Score
Normal	0.78	0.80	0.78	0.79
Benign	0.51	0.70	0.51	0.59
Cancer	0.79	0.59	0.79	0.68
Overall	0.68	0.70	0.68	0.68

Table 3: VGG-16 Results

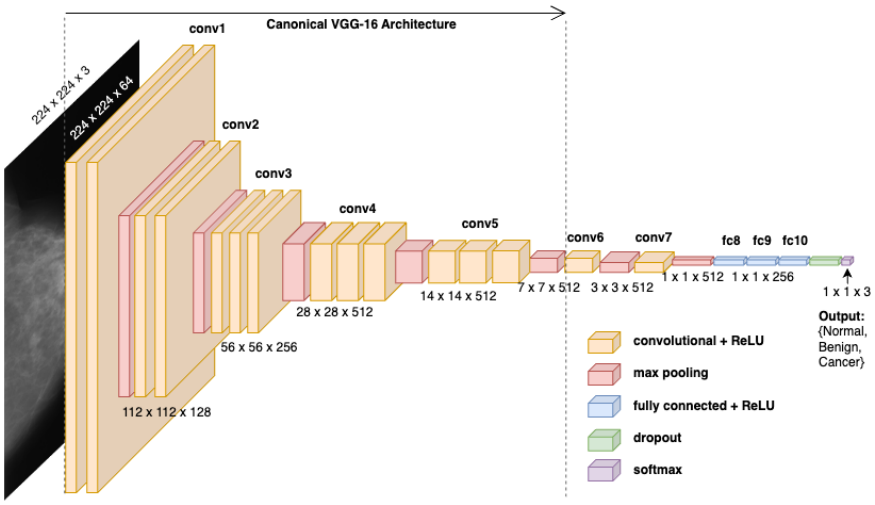


Figure 5: Network Architecture implemented with VGG-16

The VGG-16 model exhibited an overall accuracy of 68.3% after 50 epochs, surpassing the previous two architectures in performance. Notably, the VGG-16 network correctly identified 78% of the images without tumors and accurately detected tumors in 90% of benign images and 94% of malignant images, comparable to the performance of SimpleCNN. The model's precision of 0.59 for malignant images and recall of 0.51 for benign images, shown in Table 3, indicates a tendency to classify scans presenting tumors as malignant. Nonetheless, the VGG-16 architecture outperformed other models in all general evaluation metrics.

4 Summary of Model Performance

Overall, our VGG-16 implementation achieved the highest overall accuracy at 68.3% after 50 epochs of 200 steps with batch size 20. To further characterize performance, confusion matrix analyses were conducted on the three classifiers using a test set of 1936 images, shown in Figure 7. The comparative analysis of the VGG-16, SimpleCNN, and ResNet models for classifying tumor images reveals that the VGG-16 model achieved the highest accuracy (78%) in classifying images without tumors, outperforming the SimpleCNN (70%) and ResNet (56%) models. However, the SimpleCNN model showed the highest accuracy in classifying benign images, while the ResNet model exhibited the highest accuracy in classifying malignant images, albeit with a high rate of misclassification due to its bias towards malignancy.

Further analysis indicates that the low accuracy in tumor classification across all three models stems from their difficulties in correctly classifying benign tumors. Notably, the SimpleCNN model showed a misclassification rate of 45% for benign tumors, while the ResNet and VGG-16 models showed misclassification rates of 92% and 50%, respectively.

Altogether, the experimental results presented in Table 4 highlight the superior performance of the VGG-16 model in terms of accuracy, precision, recall, and F1-score compared to the other two models. Therefore, we adopt the VGG-16 architecture for further experimentation.

Model	Accuracy	Precision	Recall	F1-Score
SimpleCNN	0.65	0.67	0.65	0.65
ResNet-50	0.52	0.61	0.52	0.45
VGG-16	0.68	0.70	0.68	0.68

Table 4: Overall Results For Each Model

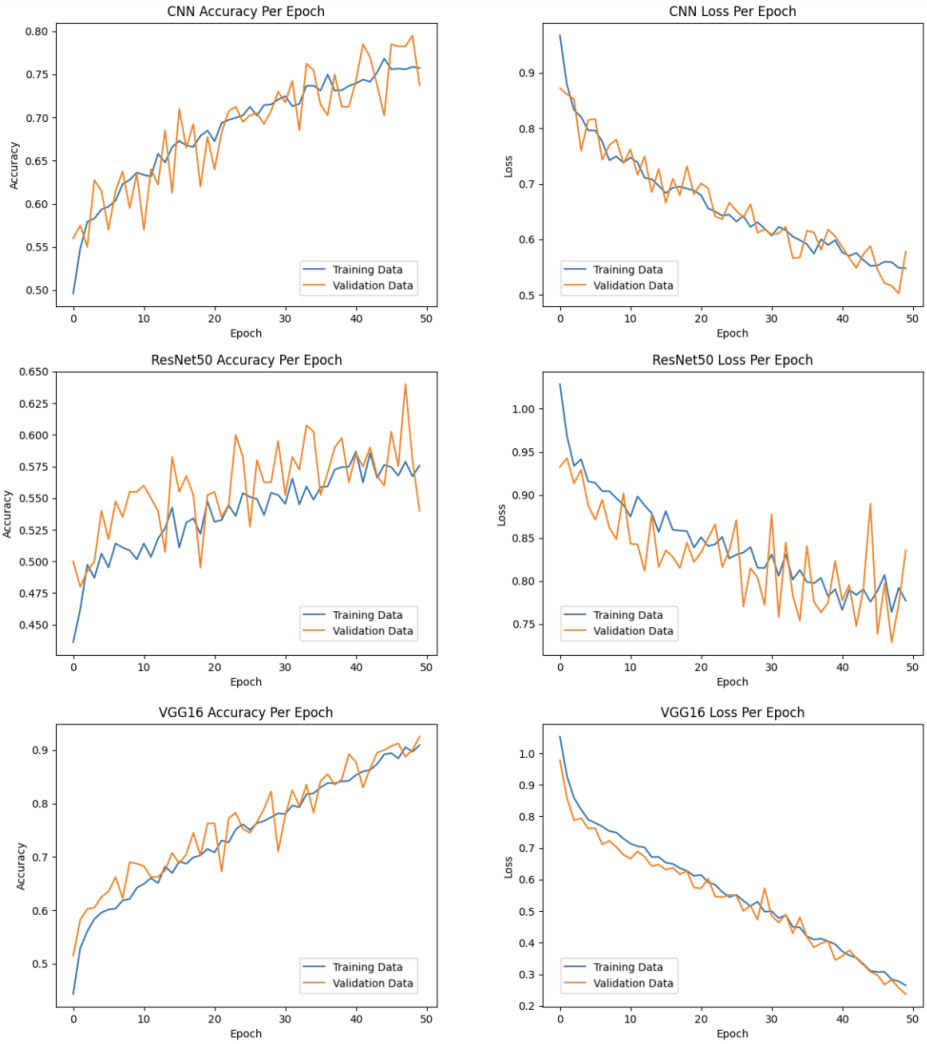


Figure 6: Accuracies and Losses on Training and Validation Sets for all 3 Models

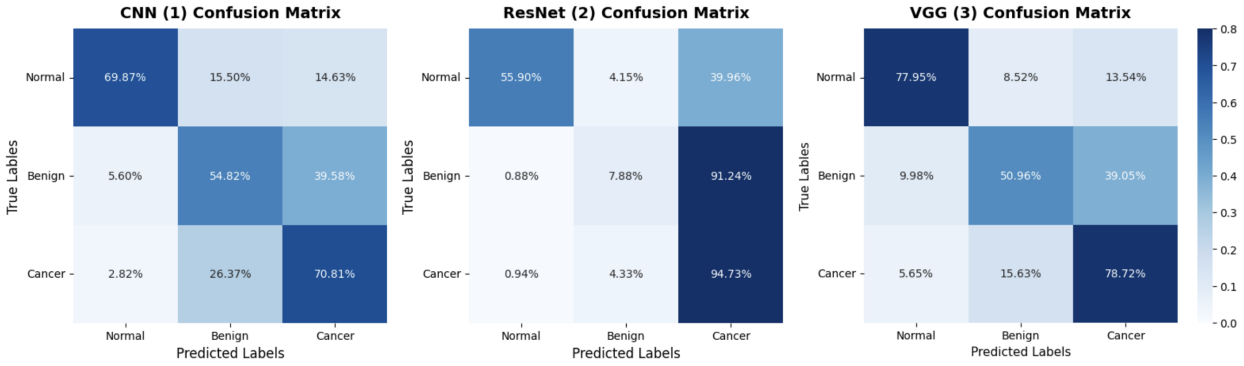
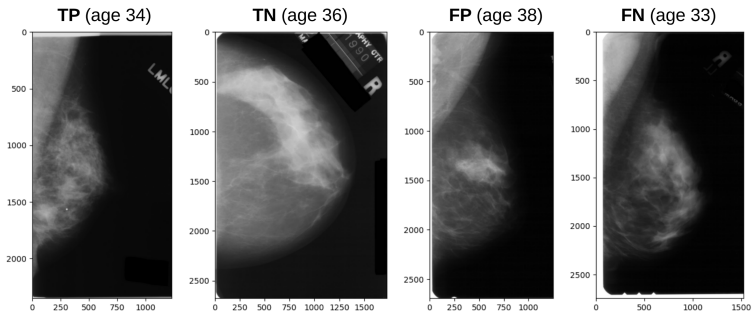
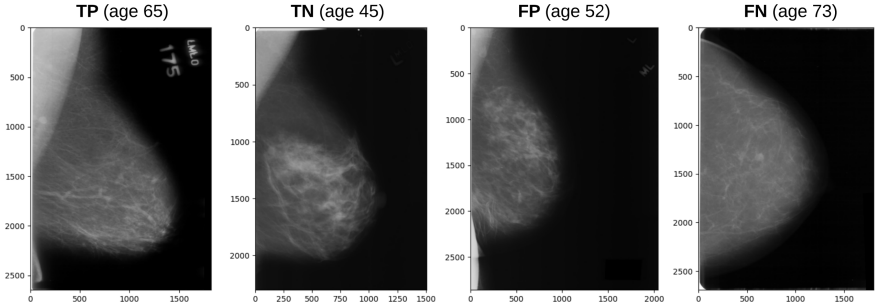


Figure 7: Confusion matrix analysis of 3-way classification with (1) CNN, (2) ResNet, and (3) VGG models on test set.



(a) Example outputs for protected age group (30-39)



(b) Example outputs for unprotected age group (40+)

Figure 8: Model outputs for 2-way classification of Malignant vs. Benign tumors across age groups

5 Experimentation Goals

With the baseline model trained, we now proceed to investigate our hypothesis after the first round of testing. Our hypothesis posits that the model performs worse for younger populations than older populations due to disproportionate data within the mini-DDSM dataset. Thus, we evaluate VGG-16 performance with regards to the age distributions of samples from the dataset.

Age Group	**30-39** (Protected)	**40-49**	**50-59**	**60-69**	**70-79**	**80-89**	**Overall**
Accuracy	0.59	0.69	0.65	0.73	0.67	0.76	0.68

Table 5: VGG-16 Accuracies By Age

Due to overwhelmingly insufficient samples from ages below 30 or above 90 leading to NaN values, we were only able to determine performance for age groups 30-89. From Table 5, it is evident that VGG-16 performance was not uniform across all age groups of interest. Specifically, accuracy scores for age groups 40-89 ranged from 65% to 76%, which are consistently higher than the 30-39 age group accuracy at 59%. These findings align with our prior hypothesis that the model's accuracy would suffer in the youngest population due to limited quantity of available training and testing data. Thus, we incorporate the results of our initial hypothesis by updating our model to utilize a greater proportion of scans and variation in images from the 30-39 age group, which we denote as the "protected" age group.

To assess impacts of age on distinction between tumor types (benign or malignant), qualitative analysis was conducted with Figure 8. In this 2-way sub-classification problem, the model outputs true positives as correctly identifying malignant masses, and true negatives are correctly-identified benign masses. A random sample of true positives, true negatives, false positives, and false negatives in the protected age group (30-39) and unprotected age groups (40+) was taken to assess potential age-related issues the model may run into. While every scan in this figure contains tumors and/or calcification in the form of white masses, their intensities appear quite different for these two groups. For example, we see more distinctive and brighter white matter in scans for the protected age group, whereas white matter appears subtle for older populations.

Thus, these appear to be reasonably distinguishable through subtle differences between the mammograms for the varying age populations. It is likely that the model encountered challenges in correctly classifying scans of younger individuals due to the insufficient amount of training data that prevented it from learning the appropriate features.

In the next section, we examine four strategies to address the issue of age-related missing data: oversampling, data augmentation, age group-specific data augmentation, and uniform sampling combined with augmentation techniques. Although we anticipate that these approaches may lead to a reduction in overall accuracy, we aim to equalize outcomes across all age groups with these approaches.

6 Data Augmentation Pipeline

To evaluate fairness of these models, we consider two auxiliary metrics. The first metric, denoted DA or Difference in Accuracy, quantifies the absolute value of the difference between the accuracy for the protected age group and the accuracy for all other age groups. This metric measures the extent to which the model's accuracy varies across different age groups. The second metric, denoted VA or Variance in Accuracy, captures the spread in accuracies across the six age groups. A smaller value of VA implies a model with more equitable predictions by age.

To obtain a comprehensive understanding of the model's performance, we compute these metrics for overall accuracy, tumor detection, and tumor classification. Tumor detection evaluates how well the model distinguishes between normal scans and scans with a tumor present, whereas tumor classification assesses the model's accuracy in classifying tumors as benign or malignant.

6.1 Oversampling

We hypothesize that one of the contributing factors to the comparatively lower accuracies observed for the protected age group is a lack of sufficient training data, resulting in an insufficiently trained model with respect to the protected age group. Thus, the first approach we utilized to enhance the accuracy of the VGG-16 model on ages 30-39 is oversampling. In this method, we sampled batches of 20 images as in the baseline model. Afterwards, we appended one additional training image from age group 30-39, ensuring that each batch includes at least one instance of a scan in the 30-39 age range. Consequently, we significantly increased the proportion of training images from the youngest age group, which we suspect will cause accuracy for this age group to increase while simultaneously reducing accuracy for other older ages as the proportion of training images from these ages is now reduced. A diagram overview of this process is given in Figure 9.

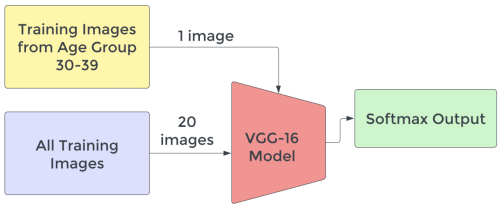


Figure 9: Diagram of Approach 1: Oversampling

6.2 Data Augmentation of All Images

Given our hypothesis that the previously implemented method may potentially result in lower accuracy rates for older age groups, we assess an alternative system aimed at augmenting the training dataset with more images across all age groups. Consequently, the second method we implemented to improve accuracy was to increase the quantity of all training images through data augmentation, which we denote generative data augmentation. To this end, we applied various augmentations such as rotation of ± 10 degrees, zooming in and out by 0.01, horizontal and vertical shifts of 0.05, shears of 0.01, and horizontal flips. A previous study of breast cancer detection also found success with horizontal flips, rotations, zoom, and intensity shifts

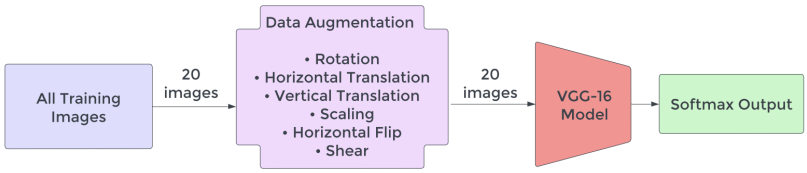


Figure 10: Diagram of Approach 2: Generative Augmentation

[11]. Examples of the data augmentations can be seen in Figure 13. We incorporated augmentations while selecting batches for training and extended the number of epochs to 70 to account for the larger training set. We anticipated that the expanded dataset would enhance the overall accuracy of the model. However, we also expected that the accuracy for the youngest age group would remain relatively low since we maintained the same image selection ratio as before. A diagram outlining this process can be seen in Figure 10.

6.3 Data Augmentation of Protected Age Group: 30-39

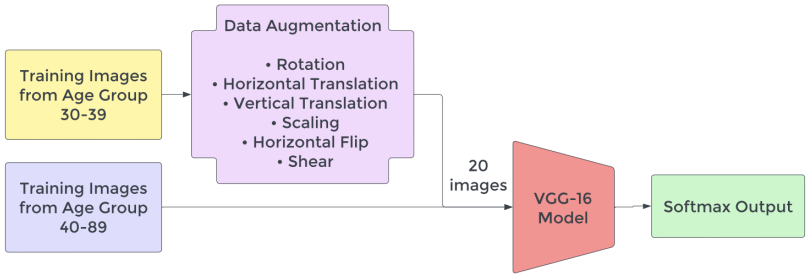


Figure 11: Diagram of Approach 3: Discriminative Augmentation

As the oversampling method reuses images and the data augmentation of all images maintains the same image selection ratio, we develop an alternative system that introduces new images with a higher representation of the protected age group, which we refer to as discriminative data augmentation. This third method combined the first two approaches to apply the same data augmentation techniques, namely rotation of 10 degrees, zooming in and out by 0.01, horizontal and vertical shifts of 0.05, and horizontal flips of the image, except now only to our protected training group. We postulate that by augmenting only the specified age range, model accuracy will increase and generalize better for test data of this age group. However, since we alter quantity proportions of images for each age group, we anticipate accuracies for unprotected age groups to decrease, although less than for the oversampling method. A diagram summarizing this process can be seen in Figure 11.

6.4 Uniform Sampling and Data Augmentation of All Images

Finally, we investigate the impact of a balanced distribution of training images across all age groups on the performance of our models. Thus, the final method we tested was uniform sampling in conjunction with data augmentation, for which we augmented all age groups to attain an equal number of images as the age group with the largest image count. This approach enabled us to achieve uniform sampling from each age group with an increased total number of images. We employed the same data augmentation strategies as the second approach and increased the number of epochs to 70 for training. We hypothesize that this method would produce the most balanced distribution and the highest overall accuracy since the model trains from a substantial number of images and an even distribution across all age groups. A diagram describing this process can be seen in Figure 12.

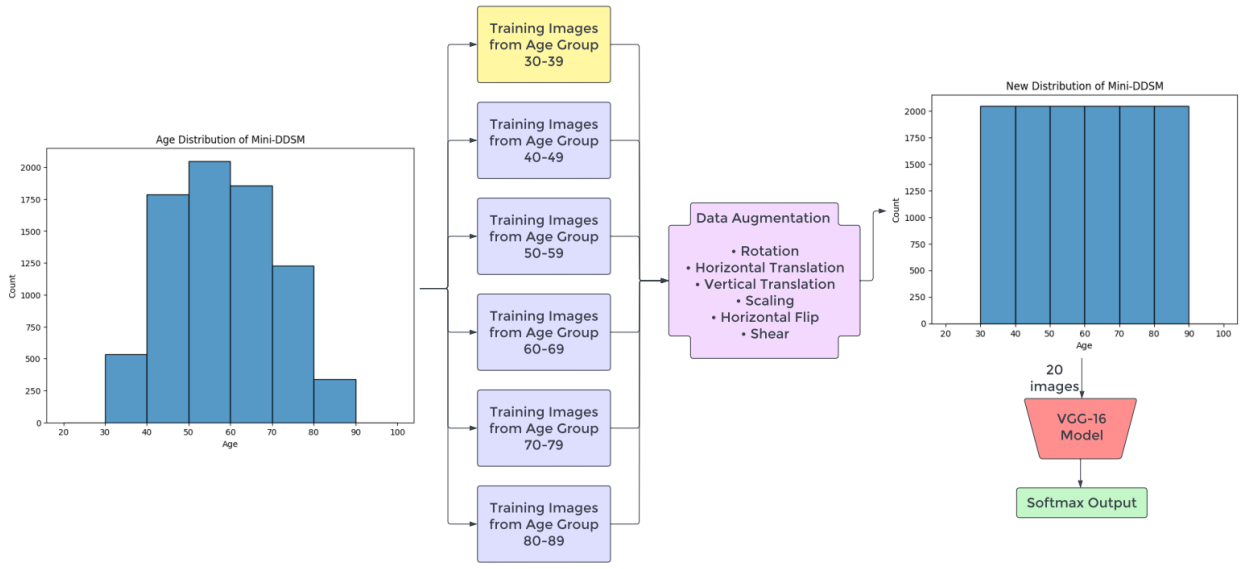


Figure 12: Diagram of Approach 4: Uniform Sampling and Augmentation

7 Results

Age Group	**30-39** (Protected)	**40-49**	**50-59**	**60-69**	**70-79**	**80-89**	**Overall**
Baseline	0.59	0.69	0.65	0.73	0.67	0.76	0.68
Method 1	0.66	0.70	0.65	0.64	0.61	0.56	0.65
Method 2	0.62	0.72	0.66	0.70	0.69	0.65	0.69
Method 3	0.65	0.70	0.69	0.66	0.61	0.65	0.67
Method 4	0.63	0.75	0.62	0.70	0.70	0.68	0.68

Table 6: Comparisons of VGG-16 Accuracies By Age

Model	**Accuracy**	**DA**	**VA**
Baseline	68.3	10.1	32.1
Method 1	65.1	1.3	20.7
Method 2	68.5	7.3	13.1
Method 3	67.1	1.7	8.4
Method 4	68.4	6.3	20.2

(a) Overall Results

Model	**Accuracy**	**DA**	**VA**
Baseline	87.9	4.1	5.4
Method 1	89.4	4.6	7.8
Method 2	89.0	5.2	11.1
Method 3	87.4	4.6	11.9
Method 4	89.4	2.6	10.5

(b) Results for Tumor Detection

Model	**Accuracy**	**DA**	**VA**
Baseline	64.3	3.2	38.4
Method 1	63.0	11.8	48.6
Method 2	63.1	0.6	10.6
Method 3	60.6	7.0	24.1
Method 4	62.3	0.6	29.6

(c) Results for Tumor Classification

Table 7: Fairness Metrics Across Methods

Based on the values presented in Table 6, our findings indicate that all augmentation techniques exhibited a positive impact on the accuracy of the protected age group. In particular, Method 1, which involved oversampling of the protected group, demonstrated the most significant improvement in accuracy for the protected age group, achieving 66% accuracy, a 7 percentage point increase from the baseline. However, this method displayed the lowest overall accuracy, potentially due to the low 56% accuracy for the 80-89 age group. On the other hand, Method 2, which applied generative data augmentation on all images, exhibited the highest overall accuracy, though only marginally higher than the baseline and Method 4 (uniform sampling). Further analysis and discussion of the results are provided in Section 8.

Table 7 presents the fairness metrics for each of the evaluated methods. Overall, we note that all methods achieved more equitable diagnoses than the baseline, as all models received lower values for both the difference in accuracy (DA) and the variance in accuracy (VA). The lower DA suggests that all models improved accuracy for the protected group while minimizing the loss for the rest of the population, since the difference in accuracy between the protected group and the remaining population was not greater than that of the baseline for any method. Furthermore, the VA column shows that these methods reduced the variance of accuracies across all age groups, indicating reduced variability and more balanced predictions across all age groups.

Interestingly, the positive results observed in the overall fairness metrics did not extend to the tumor detection or tumor classification tasks. Specifically, none of our approaches produced a smaller variance than the baseline for tumor detection across all age groups, although Method 4 exhibited lower DA than the baseline and higher overall accuracy. On the other hand, for tumor classification, all methods demonstrated lower accuracies than the baseline. However, Methods 2 and 4 exhibited a more equitable difference in accuracy between the protected age group and other age groups, while Methods 2, 3, and 4 demonstrated smaller variability among their accuracies across all age groups. These findings are discussed in greater detail in the following section.

8 Discussion and Analysis

8.1 Method Analysis

In the first method of oversampling, our hypothesis suggested that the accuracy for the protected age group would improve at the cost of overall accuracy. The findings presented in Table 6 support this hypothesis, as this method led to a higher accuracy in the protected age group but a lower overall accuracy than the baseline. This outcome is not surprising, as oversampling from the protected age group enhances the model's understanding of target scans. However, the model may begin to fit excessively to the protected age group, resulting in lower overall accuracy. Nonetheless, this model exhibited a lower DA and VA than the baseline, revealing that despite the decrease in overall accuracy, the overall results were fairer than the baseline. This is likely because oversampling results in a more balanced distribution among the various age groups and generates more comparable accuracies.

The second method, generative data augmentation, yielded a slightly higher accuracy than the baseline, along with lower values for DA and VA metrics. Our hypothesis was supported by the results, as the overall accuracy and the accuracy for the protected age group both increased, though to a lesser extent than observed in Method 1. This can be attributed to the fact that data augmentation provides more images for the model to train on, which reduces the risk of overfitting to the training data, but may not be as effective in improving accuracy for the protected age group as methods uniquely targetting this group. Even so, the decrease in VA implies that the model produced more balanced predictions across all age groups, evidence that the model learned sufficient features to accurately classify tumors across all age groups.

We theorized that the third method of age-specific data augmentation would result in a decrease in overall accuracies at the expense of the protected age group. The results, as shown in Table 6, support this hypothesis, although with a higher overall accuracy than Method 1. The DA was also much lower than the baseline and the second lowest overall, indicating that this method was effective in balancing accuracy across all ages. This approach may have worked well since augmenting the images allowed the model to generalize

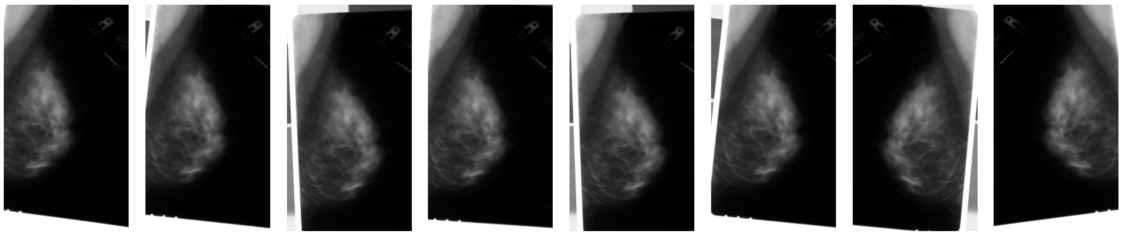


Figure 13: Examples of data augmentation

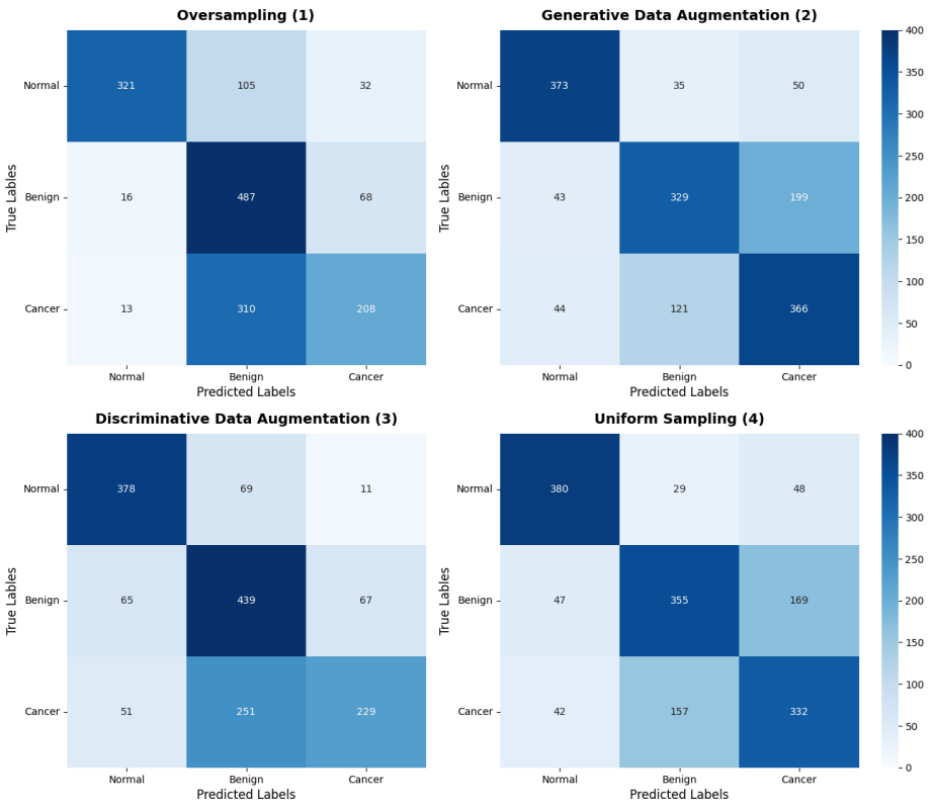


Figure 14: Confusion Matrices for Augmentation Approaches

better without overfitting to the training data. Additionally, augmenting only the protected age group helped to maintain the proportions of the dataset, resulting in a smaller loss in overall accuracy compared to Method 1. The improvement in generalization explains the decrease in VA, as the model was likely more accurate in predicting images of other classes due to the more diverse training data.

We predicted the fourth method would achieve the highest overall accuracy and the best distribution of accuracies over ages. However, our results did not meet the expected outcomes. While this approach did attain a slightly better accuracy than the baseline and a lower DA, indicating a more uniform accuracy across age groups, it did not attain the best DA among all the methods. The model's accuracy on the protected group improved, possibly due to the augmentation of all images, which helped the model generalize better, or to the increase in the number of training epochs. Nevertheless, this method did not result in the best VA and produced a relatively high spread of accuracies across age groups. This unexpected outcome may be explained by the excess amount of new data added to the original dataset through the augmentation process, causing the model to miss some patterns that existed in the original dataset and resulting in a relatively higher overall VA.

8.2 Data Augmentation Techniques

As discussed in Section 6.2 and shown in Figure 13, we applied various augmentations to increase the number of training samples and enhance the generalization ability of our model. The augmentations involved rotation of ± 10 degrees, scaling by 0.01, horizontal and vertical shifts of 0.05, shears of 0.01, and horizontal flips. The selection of these augmentations was based on a combination of qualitative analysis and experimentation with various scans. However, some of these transformations may have had negative effects on our results. As highlighted in Section 1, physical differences in benign and malignant tumors include the shape of the irregularity. Thus, certain augmentations such as shearing or scaling may have caused the model to misinterpret benign examples as malignant or vice versa, potentially leading to declines in accuracy

for tumor classification from the baseline. These questions warrant future exploration to identify the most suitable set of augmentations for breast cancer scans.

8.3 Age and Class Distribution

Upon further examination of the class distributions by age, we discovered that the protected age group of 30-39 contained a considerably different class distribution than the other age groups. Specifically, 19% of images in the protected age group were normal, 62% were benign, and 19% were malignant. In contrast, the rest of the data had a class distribution of 32% normal, 32% benign, and 36% malignant. Consequently, data augmentation techniques which focused on the protected group not only increased the amount of training data for this group but also increased the proportion of training data labeled as benign. This observation sheds light on some of the results presented in Table 7. Specifically, for tumor detection, Methods 1, 2, and 4, all of which increase the proportion of the protected group through oversampling and/or data augmentation, resulted in slightly higher accuracies than the baseline. Method 3, which used data augmentation on the entire training set and thus did not overfit to benign images to the same extent, displayed similar accuracy to the baseline. On the other hand, for tumor classification, all methods yielded lower accuracy due to the increased training on benign scans as opposed to malignant scans. Surprisingly, Method 3 displayed the lowest accuracy for tumor classification, although this could potentially be attributed to the aforementioned impacts of augmentations on tumor misclassification. As shown in Figure 14, each method rarely classified benign or malignant images as normal, but misclassification rates of malignant tumors as benign were high. Therefore, our findings highlight the importance of considering class distributions when implementing data augmentation techniques for age-based classification.

8.4 Limitations

We present some additional limitations to these results. One such limitation is the unknown scalability of our findings to more complex models. While our baseline model was able to reveal the impacts of oversampling and data augmentation, it is possible that more advanced models may respond differently to these methods. Another limitation of our study was the constrained number of epochs and steps per epoch. Due to time and resource constraints, we were unable to extend our training for longer periods, which could have resulted in better accuracies. It is plausible that additional training epochs and steps per epoch may have led to improvements in model performance.

9 Conclusion and Future Work

Recent rapid progress in computer vision, in conjunction with the increasing availability of well-annotated medical datasets, presents opportunities for further advancement of new deep learning methodologies in medical applications. This relatively new approach in medicine holds promise got improving and accelerating the diagnostic process for various cancers and diseases, including breast cancer, especially when used alongside physician diagnoses.

However, one issue that has been largely overlooked in the development of methodologies is the potential impact of age distributions. To address this issue, we created a baseline VGG-16 convolutional neural network (CNN) to examine potential inaccuracies in breast cancer classification for younger populations. Our analysis revealed that accuracy for scans of patients in the 30-39 age group was lower than of older patients.

To address this disparity, we explored four potential methods for improving classification accuracy: oversampling, generative data augmentation, discriminative data augmentation, and uniform data augmentation. Our results demonstrated that all of these methods were effective in increasing accuracy for the protected age group, with varying levels of success. Moreover, some methods also led to increased overall accuracy on the baseline due to improvements in data diversity.

However, it is important to acknowledge the limitations of our study, including the potential impact of scaling as model accuracy and complexity improves and the constraint of balancing computational resources and training time. Future research may investigate the impacts of other various augmentation techniques or

apply similar methods to more complex deep learning models to assess how these results generalize to more challenging classification tasks.

10 Notes on Implementation

10.1 Code Credit

Within this section, we expound upon the custom implementations we developed and the external source code we utilized to facilitate our research. The mini-DDSM dataset was obtained from https://www.kaggle.com/datasets/cheddad/miniddsm2. In addition, some code (around 35 lines) was taken from this website to help read in the dataset, although we updated this code to fit our own implementation. We also wrote code to pre-process the data to fit our model architectures.

For baseline model development, the pretrained VGG-16 model and the pretrained ResNet-50 models were obtained from the `tensorflow.keras.applications` package. Fine-tuned layers and layers used for SimpleCNN were obtained from the `tensorflow.keras.layers` package. However, the architecture design, hyperparameter selection, and code to create the models were written independently by authors of this paper. We also independently implemented pipelines to read in images individually, determine accuracies, and store results.

For data augmentation, we utilized the `ImageDataGenerator` from `tensorflow.keras.preprocessing` to rotate, translate, and scale images. We chose the augmentation methods and wrote code to implement oversampling, data augmentation for the youngest age group, and uniform data augmentation to fit our data.

Finally, for data analysis, we used the `sklearn.metrics` package to create confusion matrices and reports on precision, recall, and F1-score. All analyses of these plots and tables were our own. Moreover, all methods to calculate accuracy and fairness by age group and by task were created independently.

10.2 Challenges

In this section, we describe some of the challenges we faced while implementing our systems.

When importing the dataset, we found issues with Google Colab due to excessive RAM and runtime usage. Upgrading to Colab Pro did not resolve these problems, and we were unable to resolve package dependency issues in Adroit, so we instead switched to relying on our workstation CPUs. In addition, we switched from using higher resolution PNGs (45 GB) to lower resolution JPEGs (5 GB) to aid with RAM usage and better runtimes. This may have some impact on the quality of the results, as higher resolution images may have resulted in higher accuracy.

When creating our baseline models, it was very challenging to identify optimal hyperparameters and model architectures, as each model took 6-8 hours to run, even with smaller epoch lengths. In addition, given that we attempted to execute many different model versions to find our optimum, we were required to settle with running fewer epochs and smaller architectures in order to feasibly test our hypotheses. This also may have played a significant role in the discrepancy between our accuracy and state-of-the-art values. From these two challenges, we learned immensely about the difficulties of implementing computer vision programs and big data in general, as images take up a lot of space and CV models can have very long runtimes. As a result, we chose to focus more on the impacts of data augmentation rather than arduously improving the baseline model accuracy to match industry-level standards running on supercomputer clusters for weeks.

When attempting data augmentation, we ran into similar challenges as when creating our baseline, since testing even small changes in data augmentation would take 8 hours to evaluate. As a result, we were unable to test some other functions we implemented for adding contrast and performing gamma correction to the image, as well as cropping to a region of interest (ROI, Figure 15), since applying custom functions increased runtime to an estimated 22 hours per model. Specifically, the promise of increasing contrast to mimic contrast-enhanced mammography (CEM), which resolves equivocal findings of other detection scans but comes with higher radiation dosages, motivates future research [5]. This code is in the Github in Mini-DDSM.ipynb but was not presented in our final report due to time constraints involved.

A more interesting challenge we faced was that after applying higher intensity of augmentation to our training data (larger rotation/scale/translation ranges and vertical flips), inspired by results of another model

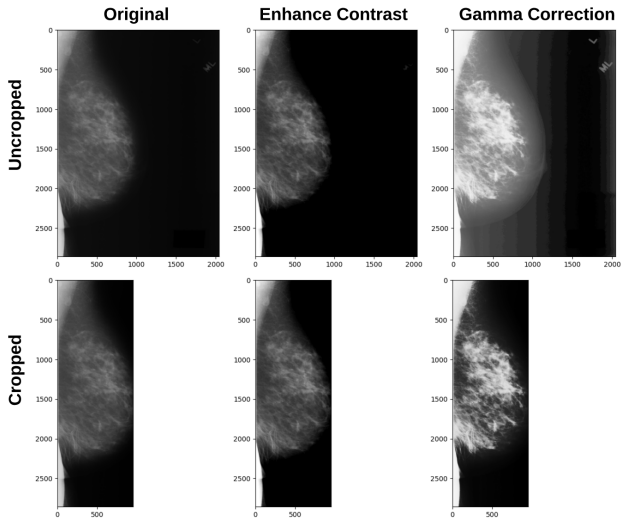


Figure 15: Examples of further image augmentation

[11], our network began to overfit to only output the benign class. We suspect this occured because over-augmenting images erases some features that were used to distinguish classes from each other, so the model overfits to one class. In general, medical scans may be treated differently than photographs taken from the natural world for classification or objection. Qualitatively, it appears that models must learn to classify whole images from patterns found in small ROIs that are often overlooked by eye (e.g. hidden tumors), ignoring distracting properties such as breast density, size, and shape. We then had to scale back the intensity of data augmention applied, but through this process, we learned that more is not always better when training models.

Lastly, selection of the best metrics to judge bias for multi-class classification presented challenges. We wrote code to compute demographic parity, equalized odds, and disparate impact, but had trouble generalizing these successfully to a three-class problem. Thus, we only used DA and VA. This code is also in the Github in the Mini-DDSM.ipynb file.

References

[1] American Cancer Society. *Dense Breast Tissue — Breast Density and Mammogram Reports*. en. Accessed on: 7 May 2023. May 2023. URL: https://www.cancer.org/cancer/types/breast-cancer/screening-tests-and-early-detection/mammograms/breast-density-and-your-mammogram-report.html.

[2] *Breast Cancer Facts & Statistics 2023*. URL: https://www.breastcancer.org/facts-statistics.

[3] *Breast cancer statistics — World Cancer Research Fund International*. en-US. URL: https://www.wcrf.org/cancer-trends/breast-cancer-statistics/.

[4] Cleveland Clinic. *Benign Tumor: Definition, Types, Causes & Management*. en. Accessed on: 7 May 2023. n.d. URL: https://my.clevelandclinic.org/health/diseases/22121-benign-tumor.

[5] Maxine Jochelson et al. "Contrast-enhanced Mammography: State of the Art". In: *Radiology* 299.1 (2021). PMID: 33650905, pp. 36–48. DOI: 10.1148/radiol.2021201948. eprint: https://doi.org/10.1148/radiol.2021201948. URL: https://doi.org/10.1148/radiol.2021201948.

[6] Constance D. Lehman et al. "National Performance Benchmarks for Modern Screening Digital Mammography: Update from the Breast Cancer Surveillance Consortium". en. In: *Radiology* 283.1 (Apr. 2017), pp. 49–58. ISSN: 0033-8419, 1527-1315. DOI: 10.1148/radiol.2016161174. URL: http://pubs.rsna.org/doi/10.1148/radiol.2016161174.

[7] Charitha Dissanayake Lekamlage et al. “Mini-DDSM: Mammography-based Automatic Age Estimation”. In: DMIP ’20 (Mar. 2021), pp. 1–6. DOI: 10.1145/3441369.3441370. URL: https://doi.org/10.1145/3441369.3441370.

[8] Andrew McGuire et al. “Effects of Age on the Detection and Management of Breast Cancer”. en. In: *Cancers* 7.2 (May 2015), pp. 908–929. ISSN: 2072-6694. DOI: 10.3390/cancers7020815. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4491673/.

[9] National Cancer Institute. *Breast Cancer Screening–Patient Version*. en. Oct. 2022. URL: https://www.cancer.gov/types/breast/patient/breast-screening-pdq.

[10] Alejandro Rodriguez-Ruiz et al. “Detection of Breast Cancer with Mammography: Effect of an Artificial Intelligence Support System”. en. In: *Radiology* 290.2 (Feb. 2019), pp. 305–314. ISSN: 1527-1315. DOI: 10.1148/radiol.2018181371. URL: https://doi.org/10.1148/radiol.2018181371.

[11] Li Shen et al. “Deep Learning to Improve Breast Cancer Detection on Screening Mammography”. en. In: *Scientific Reports* 9.1 (Aug. 2019), p. 12495. ISSN: 2045-2322. DOI: 10.1038/s41598-019-48995-4. URL: https://www.nature.com/articles/s41598-019-48995-4.