CS 254 Machine Learning
Project Group 9 - Breast Cancer Detection

1. Introduction

As one of the leaders in cancer mortality, breast cancer ranks fifth worldwide in causes for death in females [18]. In the U.S. alone, approximately 287,850 new cases of invasive breast cancer and 51,400 cases of DCIS (ductal carcinoma in situ) are diagnosed, and 43,250 died from breast cancer [1]. Moreover, it does limit itself to the female gender; males also have this sort of cancer, though the cases reported every year are nowhere near that of women cases and deaths [2]. This project aims to support the early screening and diagnosis of breast cancer through mammograms. The early detection of breast cancer contributes significantly to reducing the death rate. For this purpose, many screening methods are used, like ultrasound, screen-film mammography, magnetic resource imaging, and digital mammography [3]. Mammography is an X-ray technique that was developed specifically for breast lesion examination. Diagnosis, evaluation, and determination of the results are based on the different absorption of X-rays between different types of breast tissue [4]. A few of the injuries (small lesions) in mammograms may go undetected or be analyzed erroneously due to the quality of mammograms, the inability (experience) of the radiologists, or the limitation within the human visual system.

To bypass the issue, Radiologists use computer-aided detection/diagnosis systems (CAD) for breast cancer detection; boosting the accuracy of the CAD system can improve detection accuracy, and this will end in a better survival rate and treatment choices. CAD comprises a fundamental two-phase segmentation and classification; segmentation is an essential step in a computer-aided detection/ diagnostic (CAD) system; handcraft segmentation methods are not giving the precision recommended, are time-consuming, and are very tedious.

The tremendous progress in artificial intelligence, especially computer vision, has contributed significantly to vision systems development in various fields, such as autonomous driving [5], face recognition [6], handwriting recognition, and healthcare systems [7]. Motivated by this progress, this project utilizes deep learning to address Breast Cancer Detection. Specifically, Convolutional Neural Networks (CNN) and SVM are implemented for classification of masses, and MaskRcnn for segmentation on mammograms. The overall framework comprises three parts: 1) Data Pre-processing, 2) Classification model (CNN + SVM), and 3) Detection-segmentation based on the Mask R-CNN [8]. The overall architecture provides a flexible and efficient classification, object detection, and segmentation framework. For segmentation, after lesion extraction, it is further segmented, and the result is compared to the ground truth. The system is evaluated on CBIS-DDSM [9], a publicly available benchmark.

2. Problem Definition and Algorithm

2.1 Task Definition

Early detection of breast cancer aids in early diagnosis and treatment because the prognosis is essential for long-term survival [10], as it plays a crucial role in saving the patient's life. Previous studies have shown that early breast cancer detection prevents the spreading of malignant cells throughout the entire body, thus removing the long-term, complicated, and painful treatment process [11]. This project aims to develop a modified Convolutional Neural Network based system for early breast cancer detection.

2.2 Algorithm Definition

The overall framework consists of two parts – One is a classification network, and the other is a segmentation network. For classification, we implemented multiple CNN architectures along with transfer learning and data augmentation techniques. The main model that this work is based on is the AlexNet, which is a set of convolutional layers that extract patterns from the image, followed by fully connected layers that classify the image. Additional classifiers were implemented using machine learning techniques combined with the CNN described above. Principal Component Analysis was hybridized with Support Vector Machines, where PCA acted as another layer of feature extraction (and dimensionality reduction) from the final layer of the CNN. The PCA-transformed data was used to train the SVM classifier. As a means of comparison, a classifier using only SVM trained on the final layer of the CNN was trained as well. This will be done for the full sized CNN, as well as a truncated version to identify if overfitting occurs in the final, fully connected layers.

Regarding segmentation, we used a pretrained Mask R-CNN network for segmentation since the dataset is limited in size, with most of the default parameters. It consists of the Faster R-CNN for object detection and a fully convolutional network (FCN) for pixel-to-pixel segmentation. The Faster R-CNN uses a region proposal network to propose bounding box region candidates and then classifies these candidates into different categories. The FCN runs in parallel to perform segmentation on the regional candidates.

3. Experimental Evaluation

3.1 Methodology

3.1.1. Classification

For image classification of mammograms, we conducted multiple experiments with different CNN architectures and data manipulation techniques. A key challenge that needed to be addressed was the small sample size of the dataset. As our work is mainly focused on classifying masses, the size of the subsetted data containing masses was ~1,600 images. As we'll see shortly, data augmentation and regularization techniques were implemented as a way to address this small sample size. Also, we combined classical machine learning techniques with AlexNet and analyzed their results, where both PCA and SVM were implemented.

F1 scores, precision, and recall were calculated to evaluate performance for each model. Since the classification is for cancer diagnostics, the primary metric to maximize is recall (FP / FP + FN). False negatives in this context are classifying a patient's mass as non-malignant, when in fact it is malignant. A patient's health is most important, so false negatives are unacceptable.

A. CNN from scratch

First, we implemented a modified version of the AlexNet and trained it from scratch. The AlexNet architecture was one of the first CNNs to improve the ImageNet classification accuracy by a significant stride over traditional methodologies. The overall architecture of the used CNN is shown in Figure 1. The architecture comprises three hidden layers and two fully connected. The dataset is pre-processed and split into train-tests with the ratio of 78%-22%, respectively. The JPEG images are transformed to 256 x 256 x 1 matrices, and passed as inputs to the network. The image matrices are transformed through all the convolutional layers, the fully connected layers, and finally output as the probability of the image being malignant. Note that a sigmoid activation function is used for the final output layer, otherwise, Relu activation were applied in the convolution layers.

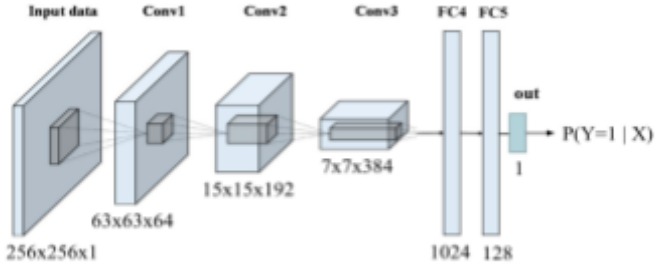


Figure 1. Custom CNN architecture used for initial classification. The architecture was based on the AlexNet where some convolutional layers (feature extractors) are followed by fully connected layers for classification (classifiers).

B. Pre-trained Alexnet

As shown in the results section, training a model from scratch strongly overfits the data. Thus, our next experiment was to use a pretrained AlexNet and fine-tune it to our small dataset. The fine-tuning is done by freezing the convolutional layers (feature extractors) and only training the fully connected ones (classifier). Note that the AlexNet was trained on the ImageNet dataset which classified samples to 1,000 different classes. So, the last layer of the AlexNet was replaced with a single neuron that has the sigmoid activation. The fine-tuned architecture is shown in Figure 2.

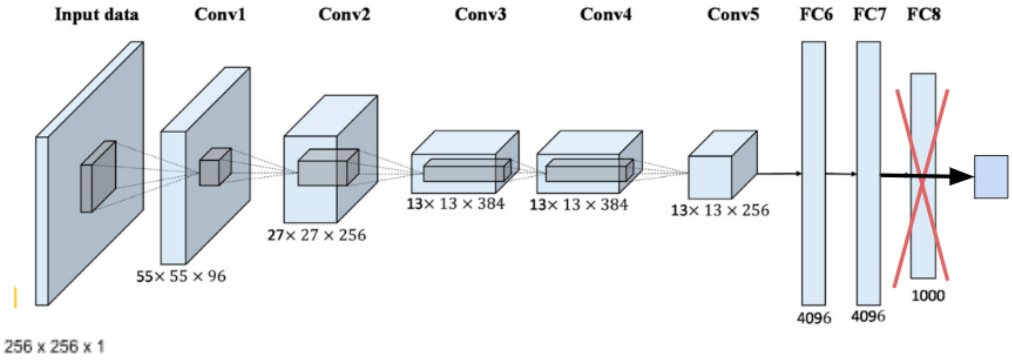


Figure 2. Fine-tuned AlexNet extends the custom CNN and uses a sigmoid activation layer for classification. Takes 256 x 256 transformed image matrix as input.

C. Data Augmentation

As overfitting was still evident even when using the above method, multiple data augmentation techniques were implemented as a way to reduce the variance in the model. Data augmentation is a method used to generate new transformed samples from the original dataset, thus, expanding the training-set size. We tried four different transformations: gaussian blurring, perspective transformation, horizontal flipping, and random rotation. A visual illustration of the transformations is presented in Figure 3.

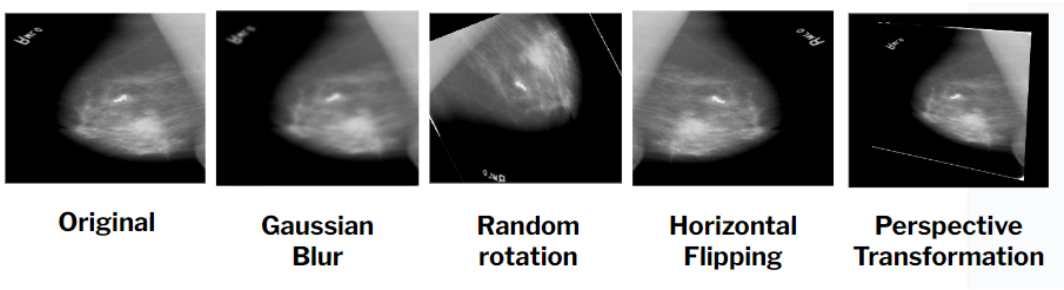


Figure 3: Data augmentation techniques applied to help solve overfitting. The augmentation was done at random, so each sample had a probability of being augmented or not, this probability was considered a hyperparameter (weak or strong).

Three experiments were carried out regarding data augmentation. We established two main methods of augmentation: weak and strong. Weak and strong augmentation correspond to the probability of augmenting the data, e.g., if it was set to 30%, then we expect 3 out of 10 samples in training to be augmented. Weak and strong augmentation had probabilities of 30% and 70%, respectively.
The three experiments were as follows:

1. Weak blurring and strong perspective transformation
2. Strong blurring and strong perspective transformation
3. Strong horizontal flipping and strong random rotations

D. SVM classifier with and without PCA as a feature extractor.

Support vector machines are an effective machine learning technique - especially when the sample size is small. For these purposes, the pretrained and modified AlexNet was utilized as a feature extractor for classification. The trained weights from the third augmentation strategy (strong horizontal flipping and random rotations) were loaded onto the network architecture. Then, the data was pushed through the trained network to effectively transform the 256 x 256 input image matrices. The transformed features in FC6 and FC7 were extracted and used to train a hybrid PCA-SVM classifier, and a simple SVM classifier. All training/test sets used a 80% training to 20% test split. All PCA-transformed data extracted 100 components from the CNN's final layer (FC7 or FC6), and SVM transformed using a polynomial kernel. This was performed using the fully modified/pretrained AlexNet using FC7 as input. Then, the classifier's were trained using a truncated version with FC7 removed - thus using FC6 as input..

The first model (Figure 4A) utilized principal component analysis combined with SVM. PCA reduced the 4096 features of the AlexNet's final fully connected layer (FC7) down to 100 transformed features, from

which SVM was trained and classified. The second model (Figure 4B) skipped dimensionality reduction and classified the 4096 features from FC7 directly (dotted arrow). Figure 4 depicts the model architecture.

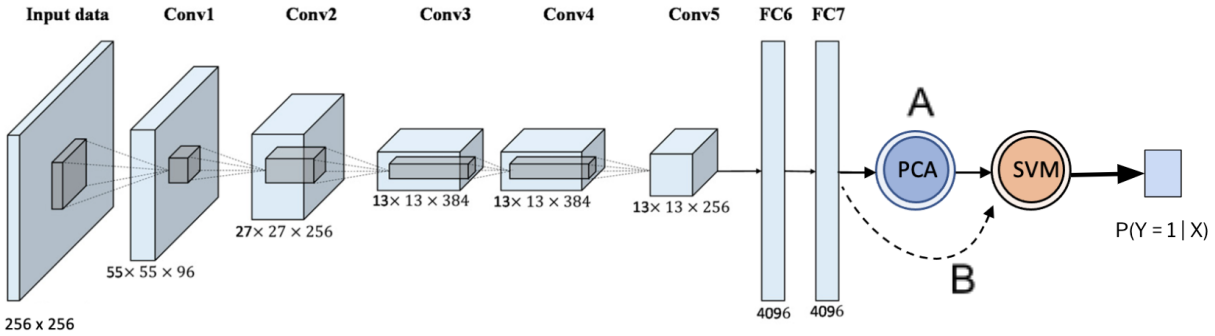


Figure 4. Classification model combining AlexNet, PCA, and SVM using FC7. (A) Model 1 uses the AlexNet and PCA as feature extractors. PCA transformed the 4096 features from FC7 into 100 principal components, from which PCA was trained. (B) Model 2 uses AlexNet's FC7 as a feature extractor to train the SVM classifier directly.

To better understand the effects of overfitting, a truncated AlexNet was created. The final fully connected layer was removed (FC7), and the same PCA and SVM models were implemented as described above. The idea was to investigate how the PCA-SVM hybrid and SVM classifiers perform when the AlexNet depth is reduced, and to see if overfitting was occurring from the final fully connected layer (FC7). The new updated architecture of the truncated AlexNet is shown Figure 5.

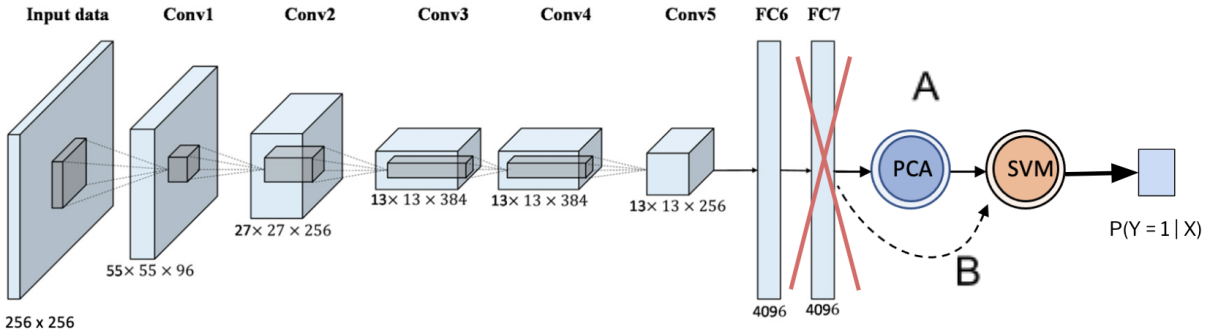


Figure 5. Classification model using the truncated AlexNet with the fully connected layer 7 removed. (A, B) The same model design was implemented as described in Figure 4, but using FC6 as the input to PCA-SVM and SVM classifiers, respectively.

3.1.2. Segmentation

Early detection of breast cancer is very important, and its accurate classification (malign/Benign) depends on the segmentation and detection of lesions; the segmentation and extraction of the region of interest (ROI) from the mammogram is challenging. We employ Mask R-CNN – a framework for simultaneous mass detection and segmentation – to segment and detect abnormalities. It includes the Faster R-CNN for object detection and a fully convolutional community (FCN) for pixel-to-pixel segmentation. The Faster R-CNN makes use of an area suggestion community to advise bounding box area applicants, after which it classifies those applicants into distinctive categories. The FCN runs in parallel to carry out

segmentation of the area applicants.

Mask R-CNN is a Faster RCNN with three output branches. The first computes the bounding box coordinates, the second computes the related classes, and the last computes the binary mask to segment the object. The overall architecture is depicted in Figure 6.

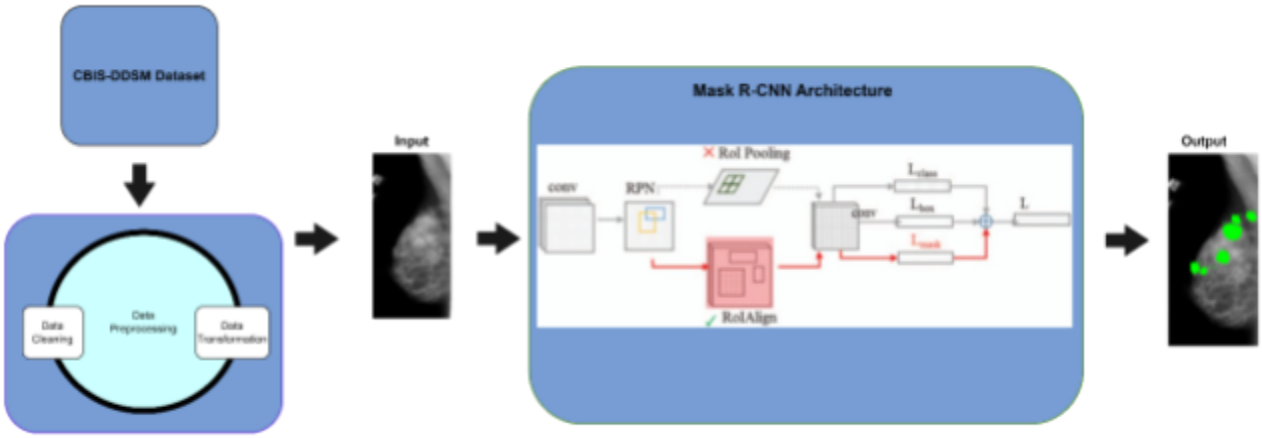


Figure 6. Mask R-CNN architecture, including input, output, and data preprocessing.

The Mask R-CNN provides a general framework for simultaneous mass detection and segmentation. It comprises the Faster R-CNN for object detection and a fully convolutional network (FCN) for pixel-to-pixel segmentation. The Faster R-CNN uses a region proposal network to propose bounding box region candidates and then classifies these candidates into different categories. The FCN runs in parallel to perform segmentation on the region candidates. The multi-task loss function of Mask R-CNN is:

$$L = L_{cls} + L_{bbox} + L_{mask}$$

L_{cls} is the classification loss, and L_{bbox} is the bounding box regression loss. The mask loss L_{mask} is the binary cross-entropy loss with a per-pixel sigmoid activation.

The Mask R-CNN training is initialized using the pre-training weights MS Coco. The ResNet101 is used as the Mask R-CNN backbone. Initially, the network was trained for both Mass and Calcification abnormalities. However, the network showed poor performance due to the varying nature of both abnormalities. Mass tumors are usually larger than calcification, which are very small calcium deposits. Therefore, we decided to train two networks – One for mass abnormalities and one for calcification abnormalities. Both networks are trained for 30 Epochs, which took around 4 hours to complete. The images are resized to (1333, 800), and random Flipping is performed on the training set. All the parameters are kept the same – (Default parameters were used).

3.2 Results

3.2.1 Classification results

Custom Classifier (training from scratch)

This model clearly overfits the training set and was not able to generalize at all, neither on the validation nor test sets. The overfitting is likely due to the small size of the subsetted data, and the lack of regularizations on the model. The training was done on 30 epoch and its results are shown in Figure 7.

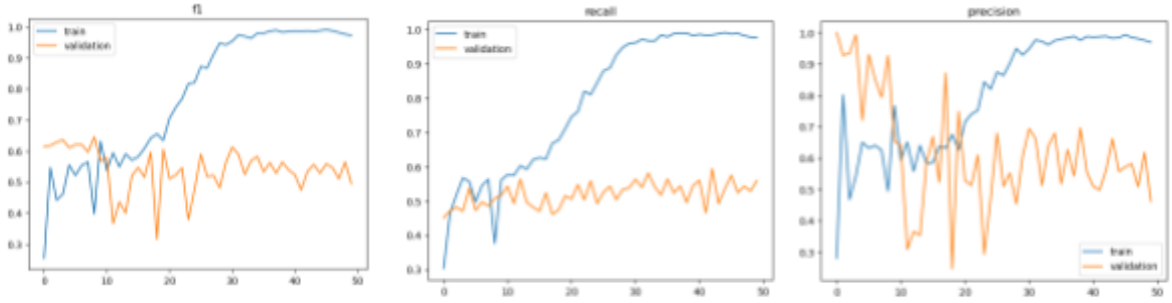


Figure 7: Custom CNN model training result. The classifier consistently yields divergency among performance metrics between training and validation sets. The model is overfitting to the noise in the training data, and is not able to generalize in the validation data.

Pre-trained AlexNet

We tried fine-tuning the last layers of the AlexNet as an initial step in dealing with overfitting. The results of such a model are presented in Figure 8, while the confusion matrix on the test set can be seen in Table 1. It is clear that overfitting was yet to be solved so we implemented different data augmentation techniques as shown in the following experiments.

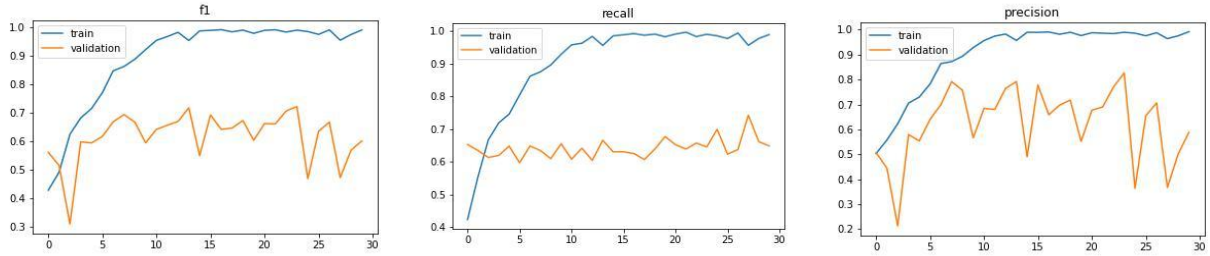


Figure 8: Fine-tuning AlexNet training results. Training and validation metrics consistently diverge from each other, indicating strong overfitting and lack of generalization past the training data.

***Table 1**: Confusion matrix of the test set using fine-tuned AlexNet.*

		Actual	
		0	**1**
Pred	**0**	119	41
	1	99	104

From the test metrics, performance is decent. F1 = 0.60, recall = 0.71, precision = 0.51, and accuracy = 0.61. In combination with training/validation plots in Figure 8, overfitting is still present. False negatives were classified infrequently, as shown by the high recall score.

Pre-trained AlexNet + augmentation 1

The first augmentation technique we used was the combination of weak blurring and strong perspective transformation. The overfitting reduced a bit but was still apparent and needed further adjustments. This experiment's training results can be seen in Figure 9, and the confusion matrix of the test set in Table 2.

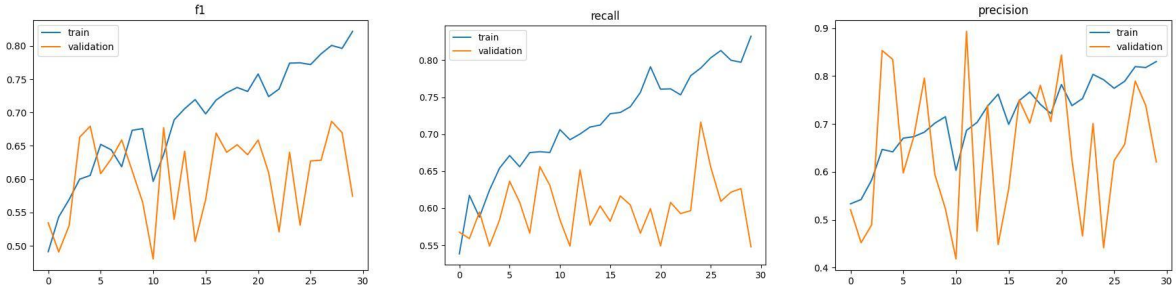


Figure 9: Training results of the first augmentation experiments using weak blurring and strong perspective transformation. Overfitting was slightly reduced from the first model, but still divergent between training and validation metrics.

***Table 2**: Test set confusion matrix from the first augmentation experiment.*

		Actual	
		0	**1**
Pred	**0**	156	55
	1	62	90

Test results from weak blurring and strong perspective transformations yield balanced but mediocre performance. F1 = 0.61, recall = 0.62, precision = 0.59, and accuracy = 0.68. Although accuracy was relatively high, this model did not achieve ideal results due to false negatives and mediocre recall. Overfitting is not as extreme here, but still unsatisfactory.

Pre-trained AlexNet + augmentation 2

The second augmentation technique we used was the combination of strong blurring and strong perspective transformation. The overfitting reduced a bit but was still apparent and needed further adjustments. This experiment's training results can be seen in Figure 10, and the confusion matrix of the test set in Table 3.

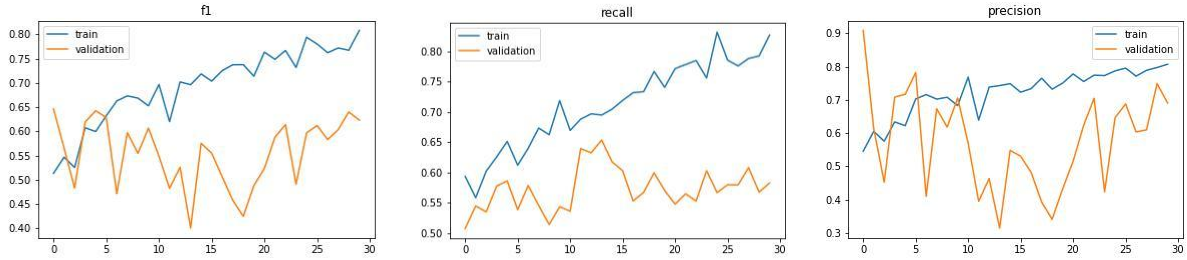


Figure 10: Training results of the second augmentation experiments using strong blurring and strong perspective transformation. Overfitting results were similar to the first augmentation.

Table 3*: Test set confusion matrix from the second augmentation experiment.*

		Actual	
		0	**1**
Pred	**0**	147	59
	1	71	86

Test results from strong blurring and strong perspective transformations yield comparable results to the previous augmentation technique. Here, F1 = 0.57, recall = 0.59, precision = 0.55, and accuracy = 0.64 - so results are slightly worse. This is somewhat unsurprising, considering the only change with this model was implementing strong blurring instead of weak blurring. This reduced overall performance when considering test statistics. Thus, it appears as though blurring may not be an effective transformation technique for this dataset.

Pre-trained AlexNet + augmentation 3

The third augmentation technique we used was the combination of strong Horizontal flipping and strong random rotations. This method was the most successful as it completely reduced the model's overfitting, and produced the highest recall on the test set. In this application specifically, we care about minimizing the false negative values, thus, the higher the recall the better. The testing recall reached 90%. The training results are shown in Figure 11, and the confusion matrix Table 4.

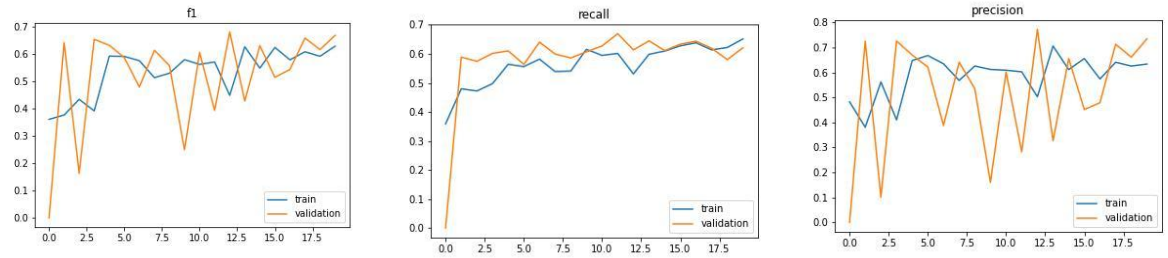


Figure 11: Training results of the third augmentation experiment using strong horizontal flipping and strong random rotations. Although the training metrics are slightly lower than previous augmented models, the validation metrics have all converged within an acceptable range to the validation. This

augmented model is far more generalizable, and has mitigated the overfitting previously observed in previously described models

***Table 4**: Test set confusion matrix from the third augmentation experiment*

		Actual	
		0	**1**
Pred	**0**	63	16
	1	155	129

Test performance metrics revealed an effective model at maximizing recall. The performance metrics are as follows: F1 = 0.60, recall = 0.90, precision = 0.45, and accuracy = 0.53. Although recall was the highest yet, the model primarily predicted the sample as malignant or 1. As such, false positives were incredibly high, yielding low precision and accuracy. However, as shown by Figure 11, the training and validation metrics converged well - indicating that overfitting was not nearly as problematic.

In an attempt to find a more balanced classifier, machine learning classifiers were implemented to compare with the AlexNet-classifier utilized above. To do this, the final layer of the modified AlexNet was modified.

Full AlexNet feature extractor with PCA-SVM hybrid and SVM classifiers

Both machine learning classifiers described below were trained using the full version of the AlexNet described in section 3.1.1D in the Methods above (Figure 4). The results summarized below display performance metrics for both classifiers.

***Table 5**: Test set confusion matrix from classifying AlexNet's FC7 using PCA-SVM classifier.*

PCA → SVM		Actual	
		0	**1**
Pred	**0**	118	54
	1	100	91

PCA-SVM hybrid test results shown here are classified by the model described in Figure 4A. F1 scores, precision, recall, and accuracy are 0.54, 0.48, 0.63, and 0.58 respectively. These results indicated reduced performance in comparison to the fine-tuned AlexNet's test performance shown in Table 1. Next, SVM classified FC7 without the reduction of features from PCA.

***Table 6**. Test set confusion matrix from classifying AlexNet's FC7 using just SVM.*

SVM		Actual	
		0	**1**
Pred	**0**	133	66
	1	85	79

SVM test results shown here are classified by the model described in Figure 4B. F1 scores, precision, recall, and accuracy are 0.51, 0.48, 0.55, and .58 respectively. Although accuracy is exactly the same as the PCA-SVM hybrid, this model has reduced recall and is worse at classifying the mammograms.

There is not much difference between the performance of the two models. The recall of the PCA-SVM hybrid is improved by ~0.08, and thus produces fewer false negatives. The reduction in features from 4096 to 100 did not change SVM's ability to classify when comparing these two models. It appears as though much of FC7 is uninformative, and as a whole does not provide improvement in capturing variance of the data. As such, FC7 is removed and both classifiers are integrated back into the network using FC6 as input.

Truncated AlexNet feature extractor with PCA-SVM hybrid and SVM classifiers

The SVM variants were trained using the truncated version of the AlexNet described in section 3.1.1D in the Methods above (Figure 5). The results summarized below display performance metrics for test sets on PCA-SVM hybrid and SVM classifiers.

***Table 7**: Test set confusion matrix from classifying AlexNet's FC6 using PCA + SVM.*

PCA → SVM		Actual	
		0	1
Pred	0	107	86
	1	111	59

PCA-SVM hybrid test results shown here are classified by the model described in Figure 5A. F1 scores, precision, recall, and accuracy are 0.37, 0.34, 0.40, and 0.45 respectively. Using FC6 from the truncated AlexNet, performance severely dropped in each metric. This is unsurprising, because there is intuitively less utility in reducing feature dimensions on an already truncated network. Using PCA on the reduced network loses too much variation in the data, which causes the SVM classifier to perform far worse than all previous methods. In contrast, using SVM independent of PCA yields improved results, depicted below in Table 8.

***Table 8**: Test set confusion matrix from classifying AlexNet's FC6 using just SVM.*

SVM		Actual	
		0	1
Pred	0	141	42
	1	77	103

SVM classification test results yielded from the model described in Figure 5B. F1 score, precision, recall, and accuracy are 0.63, 0.57, 0.71, and 0.67 respectively. As a whole, this model is the best performer. SVM classifier trained with FC6 of the truncated AlexNet improves f1 score, precision, recall, and accuracy compared to previous machine learning implementations discussed. However, the recall of this model did not reach the recall achieved from the AlexNet classifier described in 3.2.1E (Figure 9), which achieved 0.90 recall. This SVM model performed better than the previous SVM using FC7, further implying that the final fully connected layer of the AlexNet may cause overfitting of the training data.

3.2.2 Segmentation Results

The two models are trained for Mass and Calcification abnormalities, respectively. The results are not very satisfactory. This may be because the dataset is small. The nature of the dataset is also challenging. The results are shown below in Figure 12.

Mass Model Results

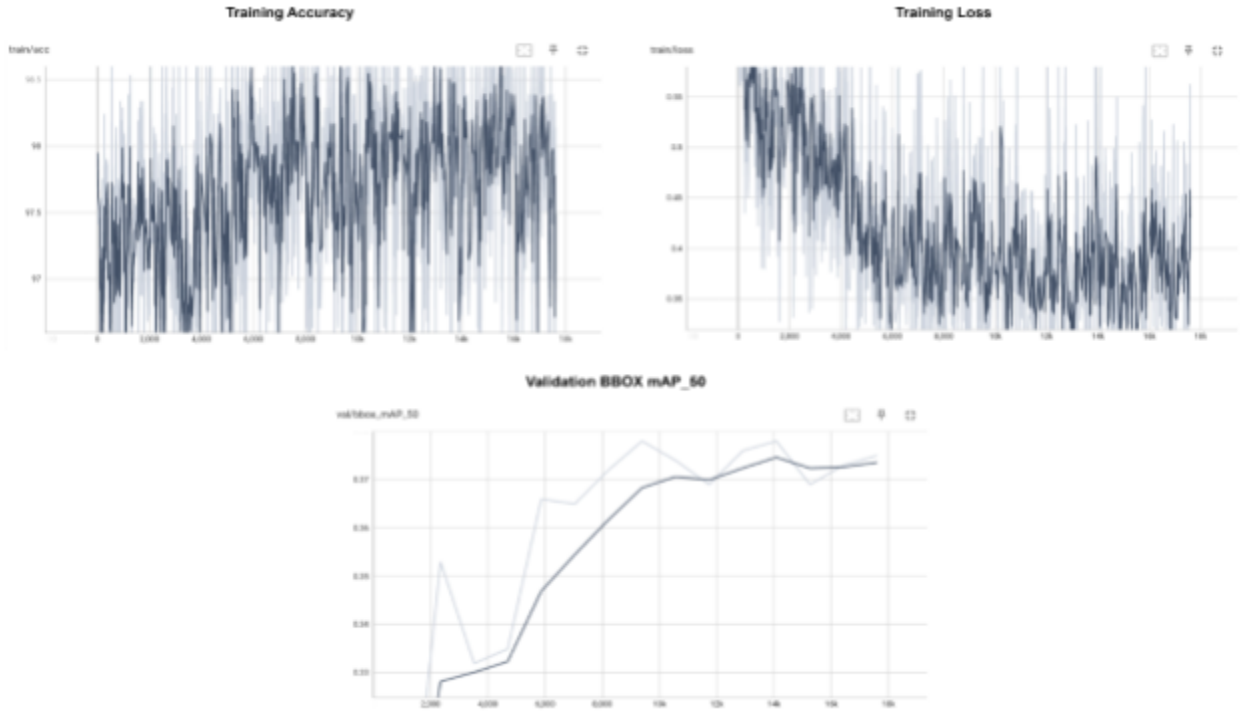


Figure 12. Training metrics evaluated from training Mass and Calcification abnormalities, respectively. Both training accuracy and training loss reach at around 8,000 steps.

The BBOX mAP_50 is the Mean Average Precision (mAP) with a threshold of 50. The mAP is a combination of precision (Out of all the positive predictions made by the model, how many are positive) and recall (Out of all the actual positive instances, how many did the model correctly identify as positive). After calculating the precision and recall values, those values are plotted to create the precision-recall curve. The area under that curve is average precision (A.P.), and mAP is calculated by calculating the mean of the A.P.s across all classes. The precision-recall curve is calculated at different thresholds; here, we used the threshold of 50. Below are the visualization results from the Mass model.

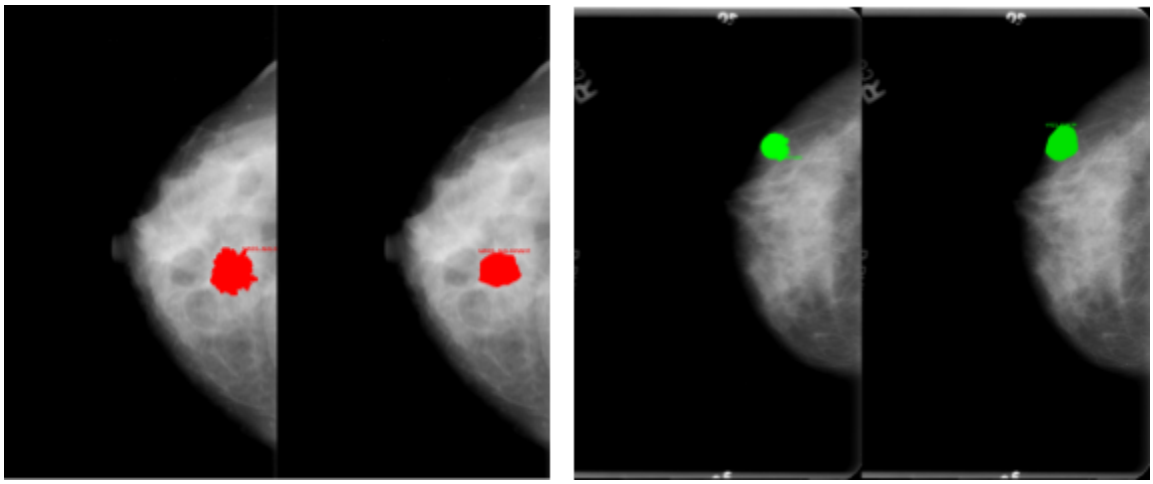


Figure 13. Results from the model. The left side of the images are original annotations of the abnormalities, and the right side is the model predictions. The green shows the Benign Masses, and the red shows the Malignant Masses. In the above examples, the model seems to be capturing abnormalities. However, there exist cases where the model captured extra abnormalities which do not exist. The model also captured F.P., F.N. Below are a few examples.

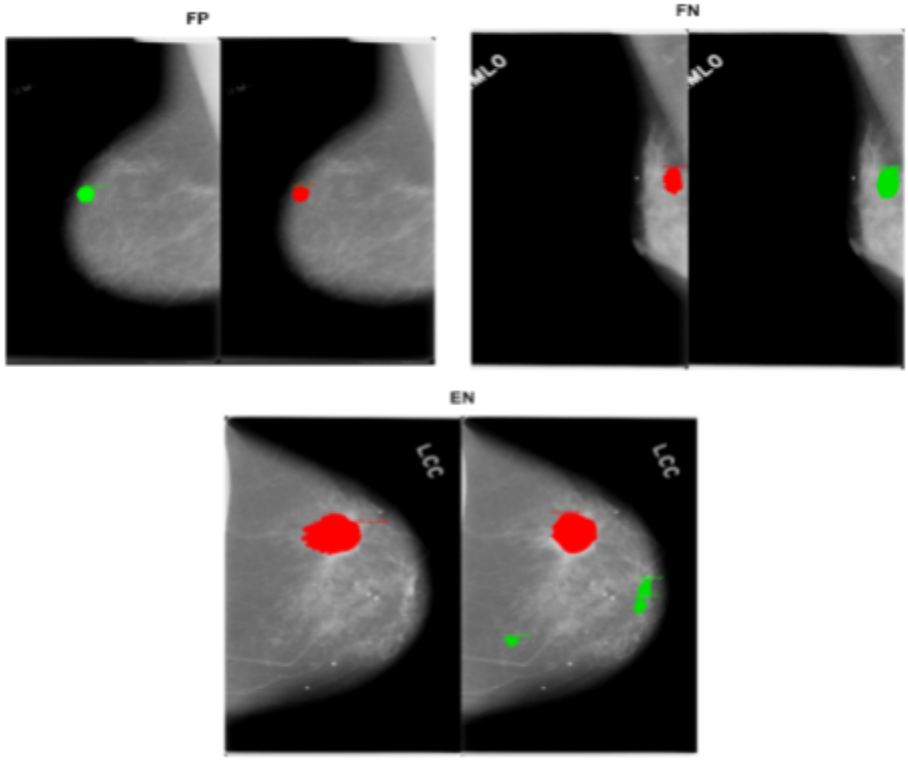


Figure 14. The errors from the model. FP represents False Positive, FN represents False Negative, and EN represents Extra Negative. The left side shows the original annotations and the right side shows the models predicted annotations.

Calcification Model Results

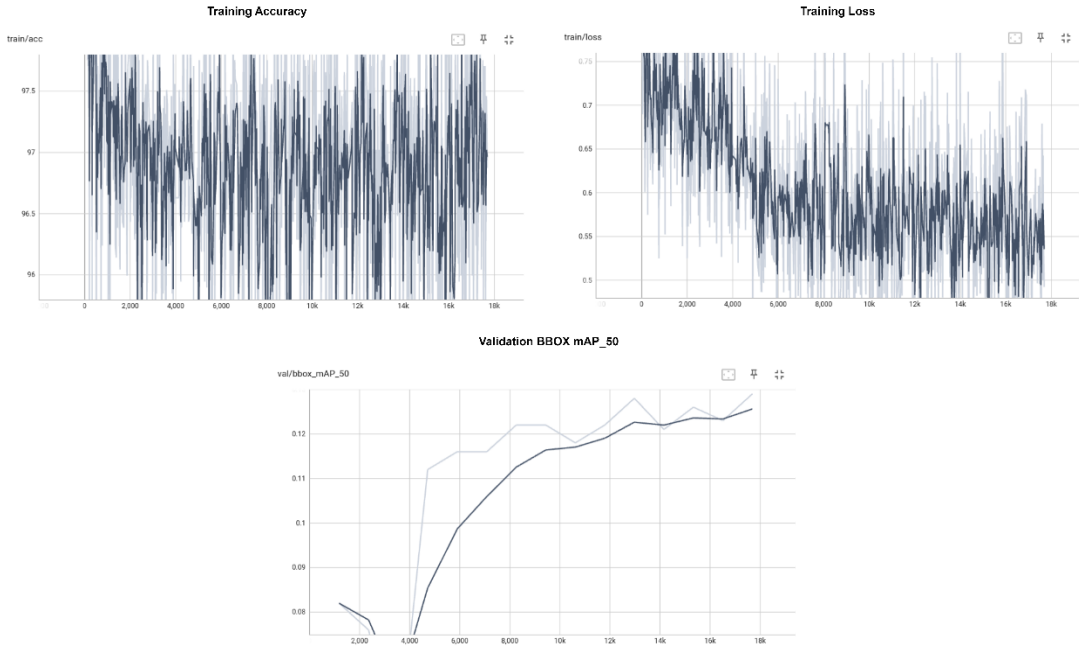


Figure 15. The calcification model exhibited poor performance in terms of BBOX mAP_50. We can see that the model is overfitting. It is performing well in training accuracy, but it is performing very poorly in validation. There are a few reasons behind this. However, before pointing out those reasons, Figures 16 and 17 visualize the results for a better understanding of those reasons.

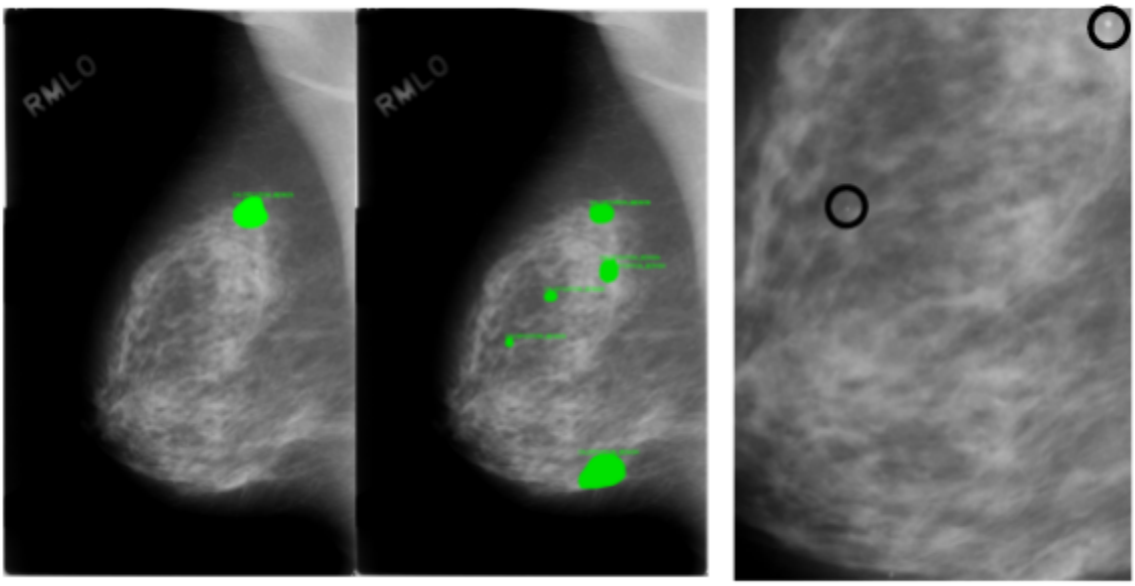


Figure 16. It can be observed from the diagram that the model is capturing the extra abnormalities which are not originally marked. However, there exists a visible pattern of calcification which appears as though it should have been marked.

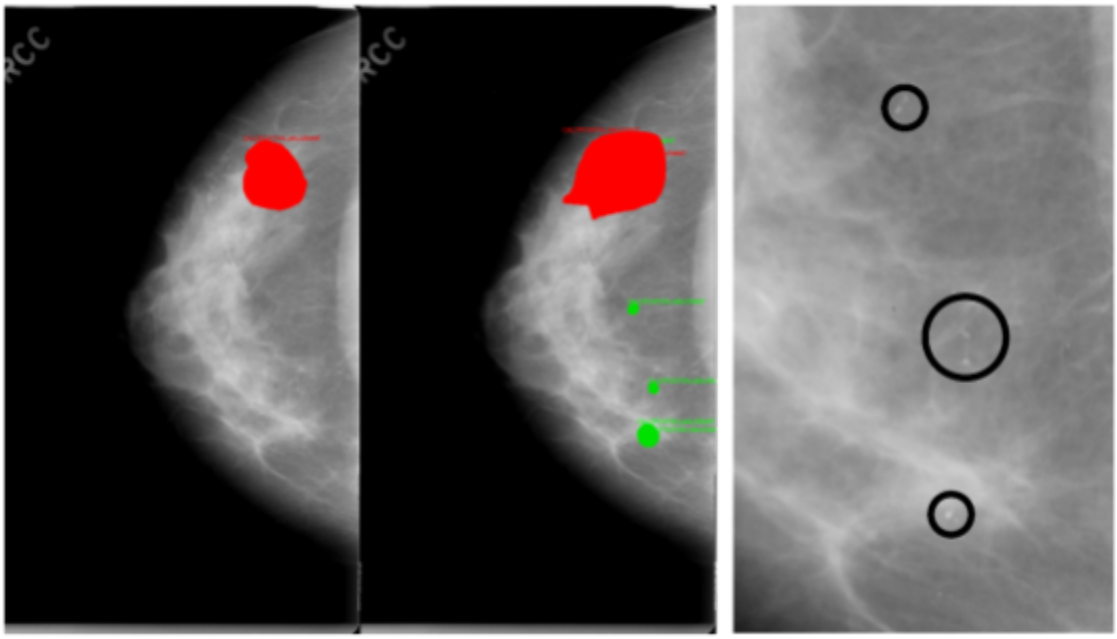


Figure 17. Similarly, it is observed that the calcifications which are not marked in the original annotations. The model is capturing those extra annotations.Model capturing the calcifications which are not marked in the dataset. This impacts the precision and recall as those captured calcifications increase the predicted positive or negative observations but do not increase the correctly predicted positive or negative observations. This way precision goes down. Similarly, this impacts the recall. As mAP is the average of the precision values at different recall levels for all classes, the mAP will also go down if the precision goes down. This is one of the challenges of using precision-based metrics.

Another reason for the weak performance of the model is the significant portion of the background. The following example illustrates it in Figure 18.

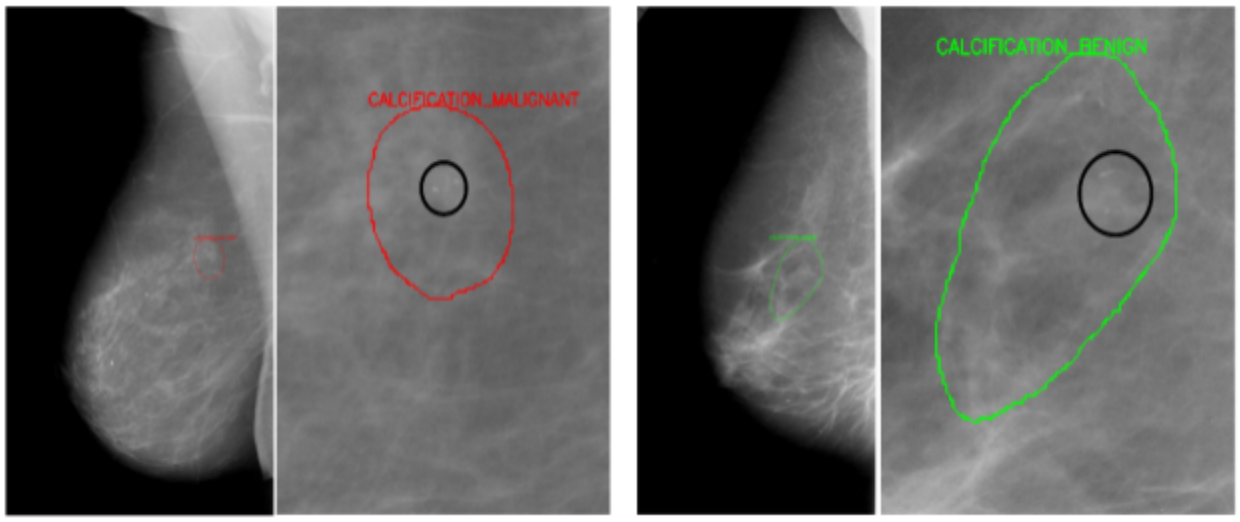


Figure 18. It is observed from the image that the annotations include a major portion of the background. In image segmentation, annotation quality is crucial to the model's performance. When an

object is annotated, the goal is to cover the exact extent of the object, excluding as much of the background as possible. If too much background is included in the annotations, it can significantly impact the model performance and metrics like Mean Average Precision (mAP). Including a significant portion of background leads to the model predicting a higher background area around the objects (as it has been trained to do so). These additional background areas in the predicted segmentation masks would be considered false positives, which would decrease the precision of the model.

3.3 Discussion

It was clearly shown how transfer learning allows us to leverage powerful pretrained models and apply them for a specific task. With the right adjustments one can produce reliable and highly accurate models, this was shown in the augmentation techniques we used. Data augmentation proved to be an efficient method that reduces a model's variance and increases its generalizability. Results indicated that blurring was ineffective at reducing overfitting. On the other hand, horizontal flipping and random rotation worked effectively for the dataset and produced the highest recall (0.90) on the test set when classified by the AlexNet. Although effective at minimizing false negatives, this model lacked in terms of accuracy and precision (0.53 and 0.45, respectively). This was likely due to some overfitting and the small sample size.

To address the sample size, the final sigmoid classification layer of the AlexNet was replaced with either a hybrid PCA-SVM or a SVM classifier. Overfitting was addressed by removing the last fully connected layer (FC7), shortening the depth of the network. In total, four machine learning classifiers were integrated into the AlexNet. The truncated AlexNet with an SVM classifier improved accuracy and precision metrics (0.67 and 0.57, respectively), but had reduced recall (0.71) in comparison to the augmented AlexNet classifier described above. The results from this model imply that the final fully connected layer (FC7) of the AlexNet may be overfitting to noise in the training data. This is demonstrated by results depicted in 3.3.1F of the Results section. Otherwise, the ensemble implementation of a modified AlexNet with SVM classifiers can be an effective method at classifying mammograms. One major limitation was the resizing of images to 256 × 256 (original images were up 3736 × 6571). In doing so, it is likely much of the nuance in images was lost. This could explain why models implemented here performed worse in comparison to some in the literature [12-17]. Also, another study found that the AlexNet performed far more poorly on the CBIS-DDSM dataset compared to the INbreast dataset, with accuracies of 0.6138 and 0.9892 respectively [20]. This could reflect some fundamental problems with the AlexNet implementation into the CBIS-DDSM dataset, which our results support.

The segmentation model provides the baseline results. It needs to be further optimized. The model seems to capture the abnormalities but also seems to be confused between classes. F.P. and F.N. cases show that the model is not well generalized. The following strategies can be tried to optimize the model's performance further.

1) Augmenting the dataset by using the Flipping, rotation to specific angles, and equalization histogram
2) Merging multiple datasets – like CBIS-DDSM and VinDr-Mammo Dataset. VinDr-Mammo supports bounding box coordinates. Therefore, Polygons from CBIS-DDSM should be converted into BBOX before using this strategy.
3) Using Resnet 50 or even Resnet 34 as the baseline instead of Resnet 101. The data is not much, and Resnet 101 makes the model overfit.

4. Related Work

Several studies have been conducted on the implementation of Breast Cancer detection and diagnosis using different methods or combinations of several algorithms to increase accuracy. S. Gcet al.[12] worked on extracting features, including variance, range, and compactness. They used SVM

classification to evaluate the performance. Their findings showed the highest variance of 95%, a range of 94%, and compactness of 86%. According to their results, SVM can be considered an appropriate breast cancer detection method.

Razali et al. [20] implemented an AlexNet on both CBIS-DDSM and INbreast to classify mammograms. They utilized rotations, shearing, and rescaling augmentation techniques on the both datasets. Interestingly, they found that performance metrics were far worse for CBIS-DDSM data compared to INbreast. The accuracy of their AlexNet was 0.6138 and 0.9892 for CBIS-DDSM and INbreast respectively.

Ahn et al. [13] presented a CNN-based approach for breast density estimation. CNN was trained to learn image features from the image patches extracted from the whole mammograms and classify them as fatty and dense class tissues. The local and global statistical features were used to train CNN. Wu et al. [14] presented the application of deep neural networks (DNN) for the classification of breast densities in D.M.s. The study comprised 20,000 screening mammograms labeled as 4 class breast densities (i.e., fatty, fibro-glandular dense, heterogeneously dense, and extremely dense). A scratch-based CNN with dense convolutional layers was used to discriminate the breast densities in the Multiview data.

Kayode et al. [15] automated an SVM classification pipeline using 234 raw mammograms from the MIAS database. They pre-processed raw mammogram images (Portable Bitmap Format) and classified them using SVM in MATLAB. Segmentation of ROIs was performed using the Otsu threshold algorithm. SVM first classifies the segmented ROI as normal/abnormal. If abnormal, the algorithm classifies it again as benign/malignant. Validation results from 78 test images (37 normal, 41 abnormal) of the first stage classification report 100% accuracy, sensitivity, specificity, PPV, and NPV. For the second round of classification (benign/malignant): sensitivity = 94.4%, specificity = 91.3%, PPV = 89.5%, NPV = 95.5%. The authors "carefully selected" the 234 raw mammograms so that the data may be biased in addition to the small sample size.

Vibha et al. [16] implemented a Decision Forest Classifier on 200 randomly selected mammograms from the MIAS dataset. Segmentation of ROIs performed before classification. They used 10-fold cross-validation, where the average is computed after 10 simulations. The classes were 151 normal, 27 "cancer," and 22 benign. They reported a classification accuracy of 90.45% averaging over 10 simulations. It is important to note this experiment was performed in 2006.

In 2021, Kumari et al. [17] implemented KNN, Random Forest (R.F.), and Naive Bayes (N.B.) to classify MIAS mammogram data. They pre-processed the images and created a feature vector using Local Binary Pattern (LBP - feature extraction technique). They further evolutionarily reduced these features by utilizing the Forest Optimized Algorithm (FOA). Evolutionarily optimized features are used as input for KNN, R.F., and N.B. classifiers. Results reported below:

Dataset	Proposed Methodology	Sensitivity (%)	Specificity (%)	Precision (%)	F1-Score (%)
MIAS	LBP+FOA+KNN	94.6	94.8	94.8	94.4
	LBP+FOA+NB	95.3	95.4	95.5	95.8
	LBP+FOA+RF	96.9	96.4	96.5	96.1

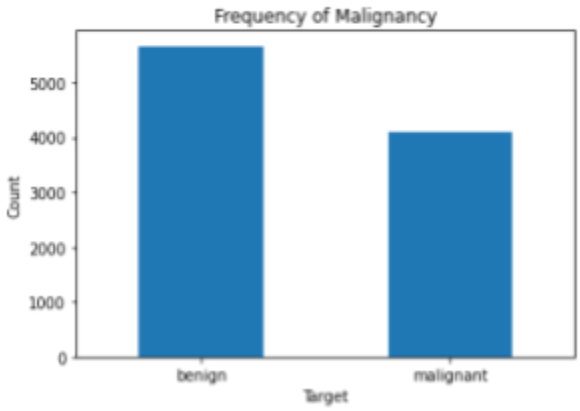
5. Code and Dataset

5.1 Dataset

Mammograms are considered a reliable and efficient source for early breast cancer detection. Mammograms are low-dose X-rays of the breast that Doctors use to look for early signs of breast cancer. Radiologists often use regular screening mammograms to find breast cancer, sometimes up to three years before it can be felt. Our dataset was a publicly available mammogram-based benchmark, CBIS-DDSM.

5.1.1 Data Pre-processing

For classification, a pre-processed version of CBIS-DDSM is used, which comprises JPEG images. The dataset is available on Kaggle [19]. This dataset contains 1566 participants and 10,239 images in total. In the downloaded dataset, there are CSV files and jpg files. The CSV files are associated with metadata containing image path, abnormality (calcification/mass), class (benign/malignant), patient I.D., location in the breast, etc. After processing, we have deduced 9764 observations to be used to analyze and train our models. The distribution of benign (n = 5668)/malignant (n = 4096) is shown in Figure 3.



The DICOM version of CBIS-DDSM is used for segmentation, available at Wiki-archive. DICOM is the standard image format when it comes to medical images. The dataset is converted to PNG format, which is usually a more programmed-oriented format. It comprises 2620 aggregated studies with scanned mammograms of benign and malignant cases with verified pathology information. The annotations are stored in Microsoft Common Objects in Context (COCO) JSON format.

The goal is to classify masses in the breast as benign or malignant and segment the individual abnormalities in those images. All the work is done based on the Python programming language and Scikit-learn, pytorch, and tensor libraries. Classes are not too imbalanced, so bias should not be a problem. As described previously, the original images were transformed from (3736 × 6571) to (256 × 256) in order to process them. For classification using the AlexNet architecture and downstream model-building, only the samples that were either benign or malignant were included. This reduced effective samples down to ~1600.

5.2 Code

All resources related to this work can be found on our Github page. It is possible to reproduce our results given that you have the dataset downloaded.

6. Conclusion

In this work, we built a system for automatic mass/calcification classification/detection system based on AlexNet and Mask R-CNN. The built system does not require handcrafted features or user intervention. AlexNet produced a classifier with a recall of 0.90, although other performance metrics were unsatisfactory. The SVM model trained on the 6-layer truncated AlexNet produced balanced results, with recall of 0.71 and accuracy of 0.68. These results were inferior to those found in the literature likely due to the reduced sample size. The Mask R-CNN showed its effectiveness to capture the abnormalities, even though the dataset was not large enough. However, it also showed that there is further need for its optimization if we consider deploying it in the real-world.

7. Bibliography

1. Giaquinto AN, Sung H, Miller KD, Kramer JL, Newman LA, Minihan A, Jemal A, Siegel RL. Breast cancer statistics, 2022. CA: A Cancer Journal for Clinicians. 2022Nov;72(6):524-41.
2. Raza SK, Sarwar SS, Syed SM, Khan NA. Classification and Segmentation of Breast Tumor Using Mask R-CNN on Mammograms.
3. Debelee,T.G,Schenker,F,rt al, "Survey of deep learning in breast cancer image analysis," Envoling Systems, vol. 11, pp. 143-163, 2020.
4. O. M. Khatib, Guidelines for the early detection and screening of breast cance, vol. 30, EMRO Technical Publications, 2006.
5. Y. huang and y. chen, "Survey of State-of-Art Autonomous Driving Technologies with Deep Learning", IEEE 20th International Conference on Software Quality Reliability and Security Companion, 2020.
6. Mei Wang and Weihong Deng, "Deep face recognition: A survey", Neurocomputing, vol. 429, pp. 215-244, 2021.
7. R Miotto, F Wang, S Wang, X Jiang and JT Dudley, "Deep learning for healthcare: review opportunities and challenges", Brief Bioinform, vol. 19, no. 6, pp. 1236-1246, 2018.
8. He K, Gkioxari G, Dollár P, Girshick R. Mask r-cnn. InProceedings of the IEEE international conference on computer vision 2017 (pp. 2961-2969).
9. Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., Moore, S., Phillips, S., Maffitt, D., Pringle, M., Tarbox, L., & Prior, F. (2013). The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. Journal of digital imaging, 26(6), 1045–1057.https://doi.org/10.1007/s10278-013- 9622-7.
10. L. Caplan, "Delay in breast cancer: implications for the stage at diagnosis and survival," Frontiers in Public Health, 2014, vol. 2, Article 87, pp. 1–6.
11. M.A. Richards, A.M. Westcombe, S.B. Love, P. Littlejohns, and A.J. Ramirez, "Influence of delay on survival in patients with breast cancer: a systematic review," The Lancet, 1999, vol. 353, no. 9159, pp. 1119-1126
12. S. Gc,R. Kasaudhan, T. K. Heo, and H.D. Choi, "Variability Measurement for Breast Cancer Classification of Mammographic Masses," in Proceedings of the 2015 Conference on research in adaptive and convergent systems (RACS), Prague, Czech Republic, 2015, pp. 177–182.
13. Ahn CK, Heo C, Jin H, Kim JH. A Novel Deep Learning-Based Approach to High Accuracy Breast Density Estimation in Digital Mammography. Proceedings of the Computer-Aided

Diagnosis; SPIE'17; February11-16, 2017; Orlando, Florida, US. 2017. https://www.spiedigitallibrary.org/conference-proceedings-of-spie/10134.toc. [CrossRef] [Google Scholar]

14. Wu N, Geras KJ, Shen Y, Su J, Kim SG, Kim E. Breast Density Classification With Deep Convolutional Neural Networks. Proceedings of the International Conference on Acoustics, Speech and Signal Processing; IEEE'18; April 15-20, 2018; Calgary, Canada. 2018. pp. 6682–6. [CrossRef] [Google Scholar]

15. Kayode, A. A., Akande, N. O., Adegun, A. A., & Adebiyi, M. O. (2019). An automated mammogram classification system using modified support vector machine. Medical devices (Auckland, N.Z.), 12, 275–284. https://doi.org/10.2147/MDER.S206973.

16. L. Vibha, G. M. Harshavardhan, K. Pranaw, P. D. Shenoy, K. R. Venugopal and L. M. Patnaik,"Statistical Classification of Mammograms Using Random Forest Classifier," 2006 Fourth International Conference on Intelligent Sensing and Information Processing, Bangalore, India, 2006, pp. 178-183, doi: 10.1109/ICISIP.2006.4286091.

17. Kumari L, K., S, J., & Rao J, R. (2021). Mammogram classification with forest optimization using machine learning algorithms. International Journal of Current Research and Review, 13(09), 136–141. https://doi.org/10.31782/ijcrr.2021.13904.

18. *Breast Cancer - Statistics*. (2023, February 23). Cancer.Net. https://www.cancer.net/cancer-types/breast-cancer/statistics

19. *CBIS-DDSM: Breast Cancer Image Dataset*. (2021, January 24). Kaggle. https://www.kaggle.com/datasets/awsaf49/cbis-ddsm-breast-cancer-image-dataset

20. Razali, N. a. M., Isa, I. S., Sulaiman, S. A., Karim, N. a. A., & Osman, M. K. (2021). *High-level Features in Deeper Deep Learning Layers for Breast Cancer Classification*. https://doi.org/10.1109/iccsce52189.2021.9530911