Localization and Classification of Lesions in Breast Tomosynthesis Scans Using 3D CNNs

Drew Clark[1], Guohao Zhang[2], Chen Bai[1], Arindam Das[3], Payman Arabshahi[1],
A. John Callegari[4], Michael Calhoun[4], and Ilya Goldberg[4]
[1]Department of Electrical and Computer Engineering, University of Washington
[2]Paul G. Allen School of Computer Science and Engineering, University of Washington
[3]Department of Electrical Engineering, Eastern Washington University
[4]Mindshare Medical, Seattle, WA

Abstract

3D digital breast tomosynthesis (DBT) is rapidly becoming the standard tool for the initial detection of breast cancer. As with all breast imaging methods, manual image analysis results in high false negative rates and extremely high false positive rates. To address this problem, we investigated the use of 3D convolutional neural networks (CNN) to automate lesion detection and classification. Localization was performed by using a 3D DenseNet classifier to output a lesion likelihood of a sub-region within a scan. Sliding the classifier over the scan with overlapping subregions produces a heatmap of lesion likelihood which are overlaid on the original scan to highlight suspicious areas. Testing localization performance on 450 samples (225 known lesions, 225 randomly sampled non-lesions) resulted in an AUC (area under the curve) of 0.973 with specificity and sensitivity of 90.2% and 94.3% respectively. Classification was performed by comparing results from a vanilla 3D CNN, 3D InceptionNet and 3D DenseNet. Testing classification performance on 225 known lesions (110 benign, 115 malignant) found the best results using the vanilla 3D CNN resulting in an AUC of 0.758 with specificity and sensitivity of 70.0% and 76.5% respectively. By adjusting the threshold to achieve >97% sensitivity, our model was able to improve specificity on our dataset from 16.4% to 30% compared to using BIRADS alone. Clinical implications of these findings are discussed.

1. Introduction

Breast cancer is the most common cancer in women worldwide[3] and the second leading cause of cancer death among women in the US [1]. Current methods to detect and classify lesions often lead to a false positive rate as high as 71.0%, representing $2.18 billion of unnecessary healthcare expenses annually in the USA[15], and immeasurable psychological trauma[14][10]. Breast scans can be difficult to diagnose due to the density of tissue in the breast, which is especially true for 2D mammograms since the image is a projection of a dense 3D space. 3D digital breast tomosynthesis (DBT) has been shown to improve specificity over 2D mammography and MRI[11] , but is relatively new to the field and lacks robust computer aided diagnosis (CAD) systems compared to 2D mammograms. DBT uses similar techniques as 2D mammography but produces a 3D image, which unlike typical 3D images, consist of high resolution slices (roughly 10 voxels/mm in the $x - y$ plane) along a slightly curved z-axis at a lower resolution (roughly 1 voxel/mm). As this technology continues to gain popularity in the field, it is desirable to develop a CAD technique using DBT scans to detect lesions earlier and to more accurately characterize lesions to reduce unnecessary medical procedures.

Our goal is to improve the accuracy and efficiency of breast cancer diagnosis using AI to automate two aspects of the diagnostic process: lesion localization and lesion classification. This paper treats the detection and classification as separate problems in two independent stages. The first stage is localization, where an entire scan is input into the localizer which produces a heatmap of lesion likelihood to be used as an aid to the radiologist. After this first stage, the radiologist may choose suspicious areas, whether detected by the heatmap or not, to be fed into the classifier. The second stage is classification, which takes a region of interest (ROI) and outputs the classification of the region containing a benign or malignant lesion. This approach keeps the radiologist in control, and allows the radiologist to combine their knowledge with the method presented in this paper to help detect and classify lesions.

2. Literature Review

Although a lot of research has been performed on developing CAD models for breast cancer imaging, the majority have used 2D mammography. A recent review paper by Harvey et al[5] discusses state of the art 2D mammography CAD systems which include Carnerior et al.[5] who achieved an AUC of 0.9, Teare et al[13] who achieved 80% specificity and 91% sensitivity, and Kim et al[8] who achieved an overall AUC of 0.903, 75.6% sensitivity and 90.2% specificity.

The major reason for a lack of research in DBT CAD is the lack of available DBT datasets due to the increased requirements of labeling 3D datasets[5]. At the time this paper was written, only two 3D CNN approaches for DBT images have been attempted to our knowledge. In 2018, Zhang et al[18] compared classification results of 2D mammography and DBT on a proprietary dataset consisting of 3018 benign and 272 malignant images. They were able to achieve an AUC of 0.7237 on the 2D dataset and 0.6632 on the 3D dataset. Also in 2018, Wichakem et al[16] developed an end-to-end 3D CNN to detect malignant lesions using a proprietary dataset consisting of 24 benign cases and 91 malignant cases to achieve 72.7% accuracy and a 0.842 F1 score on a test set of 22 instances (18 malignant, 4 benign).

Other research has been done that applies 2D CNNs to DBT images with better results achieving AUC values of up to 0.91[12]. However, classification and localization on other cancers, specifically lung cancer, have shown that better results can be achieved with 3D CNNs compared to 2D CNNs. In 2017 Kaggle's data science bowl[6], 3D lung CT scans were used for localization and classification of lung nodules with the winner using a 3D CNN[4]. Although DBT is inherently different than CT scans, it makes intuitive sense that with enough data and computational power, 3D CNNs would be able to leverage spatial relationships between slices in order to produce better predictions compared to 2D CNNs.

3. Problem Definition

Our objectives in this work are twofold. First, given a DBT scan, we attempt to detect and locate the presence of a lesion or lesions. Throughout this paper, we refer to this objective as the *localization problem*. Second, if a lesion or lesions are present, we attempt to classify the malignant or benign lesion. Throughout this paper, we refer to this objective as the *classification problem*.

Detection and localization of lesions are currently performed by trained radiologists. For localization, we attempt to create a machine learning model which can perform a task currently done by skilled professionals using the scans alone.

Radiologists are tasked with correctly classifying patients that require additional diagnostic procedures. They use their clinical judgement, taking into account information about the patient and features visible in the breast scans. The classification is considered correct if a subsequent biopsy is scored as malignant by a pathologist. Thus, we attempt to create an AI tool which can perform a task that typically requires a high level of skill.

Further details on these two problems are provided below.

3.1. Localization

For localization, our goal is to train a CNN to take an entire DBT scan as an input and output a heatmap of lesion likelihood. Each DBT scan may be of different size and resolution, so pre-processing must be performed to normalize the images to a fixed resolution and the neural network must be able to deal with inputs of varying sizes. The dataset includes labels of lesion centroids in each scan that have been identified by radiologists and clinically confirmed to be lesions, though it is possible that not all lesions in a scan have been identified. These centroids are used as positive samples (lesion) to train the localizer. Negative samples (no lesion) are obtained from areas of a scan not containing lesions.

3.2. Classification

For classification, our goal is to train a separate CNN capable of reducing the false positive rate, without increasing the false negative rate, compared to currently used BIRADS score based baseline standards. The CNN takes a ROI from the DBT scan as an input and outputs the classification of the ROI being positive (needs further intervention) or negative (needs no further intervention). As with localization, since each DBT scan can have a different resolution, all scans must first be normalized. The ROI size is predetermined so the network always receives the same input shape. The centroids are used to create the ROIs and the diagnoses are used as labels to train the classifier.

Input features include the ROI from the scan and the BIRADS score for each lesion. The BIRADS score is a standard score based on the scan that can range from 0-5. A higher BIRADS score indicates a higher probability of malignancy. This score may be used as an additional input feature since recent studies have demonstrated the benefit of including it[17]. Each lesion is labeled with a final diagnosis, either *benign* (target classification of negative) or *malignant* (target classification of positive) as determined by final clinical outcome, that is used as the ground truth labels for classification.

Table 1: Data splits

Data Overview			
Diagnosis	*Test*	*Val.*	*Train*
Benign	110	100	468
Malignant	115	80	253
Total	225	180	721

4. Data Overview

A total of 1052 DBT scans were obtained and annotated by Mindshare Medical. The dataset contains 608 unique lesions (multiple lesions may exist in a single scan), the majority of which have two separate scans to capture different views: mediolateral-oblique (MLO) and Cranial-Caudal (CC), which are perpendicular views from the side and from the top, respectively. These views contain different information about the lesion since the images have lower resolution along the curved z-axis. As such, the different views were treated as individual samples for this paper, resulting in 1126 total lesion samples.

These 1126 samples were then split into training and test sets, with 20% of the samples being used for the test set (225 samples). A validation set was created by taking 20% of the remaining training set (180 samples), leaving 721 lesion samples for the training set. Since some of the scans were different views of the same breast, or from the same patient but different breasts, it was also necessary to ensure that the test/validation sets contain different patients compared to the training set. Table 1 summarizes the data split by unique lesions, Figure 1 shows some examples of the raw data, and Figure 2 shows a sample of an ROI image of a lesion.

Using the radiologist-determined BIRADS value from the initial scan and knowledge of the final diagnosis, we are able to calculate a baseline sensitivity and specificity of the classification problem for our dataset as performed by radiologists in standard clinical practice. A caveat of our dataset is that it only contains patients who were sent for a biopsy, so although it isn't representative of the population as a whole, it does serve as a good comparison for our classifier performance. A BIRADS score of 3 usually results in a clinical recommendation for a follow up scan, which if benign, leads to unnecessary cost, time and stress for the patient. For the purpose of establishing a baseline which would allow a comparison of our work to current clinical practice, BIRADS values in the range 3-5 are deemed positive since further intervention in the form of follow up scans or biopsy are recommended. Conversely, BIRADS values less than 3 are deemed negative since no further intervention is recommended. The resulting baseline specificity and sensitivity for our dataset using BIRADS score alone are 16.4% and 97.8%, respectively. Table 2 summarizes these baseline metrics.

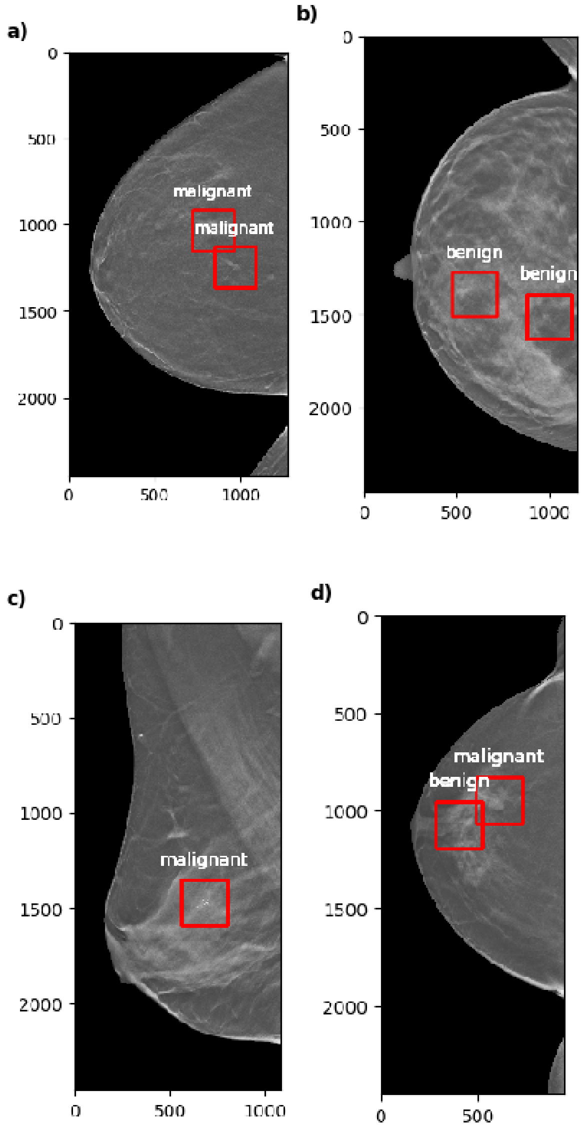

Figure 1: Data examples in test set (single slice at centroid). a) Image with two malignant lesions b) image with two benign lesions c) image with one malignant lesion d) image with one benign and one malignant lesion.

Significant challenges with the data relevant to both the localization and classification problems are discussed below:

Amount of Data: Sample size is a common constraint for machine learning in the medical industry, and especially so for breast cancer DBT scans. One issue with medical images is privacy concerns limiting the amount of data avail-

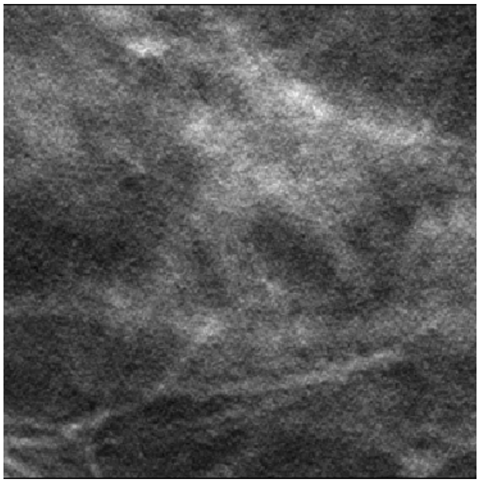

Figure 2: Example of single lesion ROI (single slice at centroid)

Table 2: Classification problem: baseline confusion matrix for our dataset using BIRADS score only.

Confusion Matrix			
	Prediction		
Actual	*Negative*	*Positive*	*Class Accuracy*
Benign	111	566	16.4%
Malignant	10	439	97.8%

able publicly. Additionally, annotating the data can be very time consuming and costly. For localization, creating new annotations is a manual process requiring radiologists to examine each image. Classification requires linking identified lesions to downstream follow-ups of the patient to determine the clinical outcomes. Data augmentation was required in an attempt to address this data scarcity issue.

Breast Density: As can be seen in Figure 1, breast density can vary significantly from patient to patient which makes both localization and classification challenging. One of our key objectives was to develop models which can locate and classify lesions in high density breast tissue.

Data Value Normalization: The dicom files contain normalization factors that we assume normalize all the files to each other, according to the dicom standard. However, the dicom scale is Hounsfield units, and the pixel values that result from the normalization factors do not match the X-ray densities expected for soft tissue on the Hounsfield scale. So we assume that the data are normalized to an absolute scale, but not in Hounsfield units.

Data Labels for Localization: In order to train the localizer model, centroid locations (x, y, z values) are provided as labels, but the size and shape of the lesion are unknown. Without ground truth masks or information on size or shape of the lesion, training of the localizer must be performed by using only a single known pixel for lesion location. Since we don't have a true representation of each lesion to compare to when producing our heatmap, this is a challenge while training. We address this by performing localization via a sliding classifier that outputs the probability of an ROI containing a lesion and use the single point to create an ROI which we know contains a lesion.

5. Preprocessing

The ROI size for both the localization and classification tasks was 2.4 cm^3, equal to (240, 240, 24) voxels. An identical ROI shape was chosen for both tasks so that similar pre-processing steps can be applied. The existing lesion samples were all 1126 known samples in our dataset. Since only a small portion of each scan is a lesion, non-existent lesion samples are easily obtained from the remainder of scans. Heavy data augmentation was used to increase the amount of samples. These steps are performed in the preprocessing stage and explained further below:

Cropping: Images were cropped to decrease the computation time and free up memory. As seen in Figure 1, many of the images contain large areas which cannot be utilized due to the nature of the scan. Labels were updated with the new centroid(s) of the lesions based off the updated origin.

Pre-cropping Existing Lesion ROIs: Since cropping is a fast process compared to normalization, the image is "pre-cropped" around each centroid to an area large enough to accommodate data augmentation (rotations and shifts) prior to normalization.

Normalization of Existing Lesion ROIs: Normalization was then performed on the cropped images. Information on the original spacing between voxels was extracted from the dicom file, and a target spacing of (0.1mm, 0.1mm, 1.0mm) was used to normalize all the images. Additionally, labels were updated with the new centroid(s) after normalization.

Augment Existing Lesion **ROIs:** Data augmentation was performed to address the low sample size issue. The classification task has a data imbalance issue between benign and malignant, which was balanced out through data augmentation. For each ROI, rotations of 90, 180, and 270 degrees were performed, followed by random shifts, rotations, and mirrors such that a benign ROI has 15 random augmentations, and a malignant ROI has 30 random augmentations. Random augmentations are constrained such that shifts are no more than 20 percent of the ROI size, and random rotations are no more than 25 degrees. In total, each benign lesion results in 19 ROIs, and each malignant lesion results in 34 ROIs. Data augmentation techniques were chosen in order to simulate what we would expect to see in reality to help our model generalize to realistic scenarios.

Final Crop for Existing Lesion ROIs: Once the augmentation was performed, this step returns the ROIs with

Table 3: Data splits after augmentation

Data Overview				
Diagnosis	*Test*	*Val.*	*Train*	*Augmented Train*
Benign	110	100	468	8892
Malignant	115	80	253	8602
No Lesion	225	180	721	17494
Lesion	225	180	721	17494

dimensions (2.4, 2.4, 2.4) cm.

Non-Existing Lesion ROIs: In order to keep an even class balance for the localizer, each scan extracts an equal number of random ROIs with no lesion as the number of ROIs containing lesions produced by that scan. The centroids are used to first pre-crop the ROI, then normalized, and finally cropped to (2.4, 2.4, 2.4) cm.

The main difference between the localizer data and the classifier data is that the latter uses only the existing lesion data with malignant or benign labels while the localizer uses all the classifier data as known lesions and an equal amount of ROIs not containing a lesion. Table 3 shows the final amount of data after augmentation and the test/validation/train splits.

6. Approach

6.1. Localization

The architecture used for the initial localization classifier is a multiple output 3D DenseNet [2] with a single input (240x240x24 ROI). Due to the low sample size, overfitting was difficult to avoid and the following changes to the model were made to help prevent it. Instead of a fully connected layer, a Global Average Pooling (GAP) layer was used as the final classifier since a GAP layer has been shown to act as a regularizer and helps to control overfitting[9]. Other methods adopted were dropout after the GAP layer during training and batch normalization after each convolutional layer. We observed best results by using a very high dropout factor (as high as 0.9) as opposed to decreasing the size of the model. We believe this to be the case due to the complexity of classification in these images requiring a high number of features to be extracted by the convolutional layers. However, by not reducing the size of the model, our network is prone to overfitting due to the low number of samples available; therefore, a high dropout factor was used to restrict the number of features available to the final classification layer during training and promote generalization. Exponential linear unit (ELU) was used as the activation function for each layer, except for the final classification which used a softmax activation. An Adam optimizer was used with accumulation to allow a larger batch size despite memory constraints. This is done by accumulating the gradients over a specific number of batches before applying the gradient, and allows a larger effective batch size when memory is limited. In our work, we adopted an effective batch size of 128 by using a batch size of 8 and accumulating the gradients over 32 batches before updating. Finally, the learning rate was adapted based on the validation set accuracy. A “leaky bucket” method was used such that if the validation accuracy decreases compared to the previous epoch, a counter is incremented, otherwise it is decremented. The learning rate was reduced by a factor of 10 whenever the counter reached 3.

6.1.1 Training and Implementation

Once the localizer is trained, it slides across a cropped scan to produce a heatmap of lesion likelihood. A stride of 20 along the x and y axes and 2 in the z axis was used, resulting in a high overlap to produce a high resolution heatmap. Since each ROI of (240, 240, 24) voxels results in a single likelihood score, the resulting heatmap is smaller than the original image and must be resized to match the dimensions of the input image. This is done by replicating the likelihood score to the entire region and averaging the scores for each pixel provided by the overlapping stride. The heatmap is then padded with zeros to account for the initial cropping and un-normalized such that the heatmap is the same size as the original image and can be overlaid and used with the original scan to help detect lesions.

6.2. Classification

For the classification task, we used the 3D DenseNet architecture described in the localizer section, as well as two additional architectures, a 3D vanilla CNN described in Table 4 and a 3D InceptionNet V2 [7] which were slightly modified from implementations already developed. Like the DenseNet, InceptionNet is a common CNN for 2D image classification and the architecture can be readily found. All models were implemented in Keras using TensorFlow.

Similar techniques were used in the classifier task as in the localizer task described above. However, prior to classification, the normalized (between 0 and 1) BIRADS score was included as a feature and concatenated with the input to the final layer prior to classification.

7. Results and Discussion

7.1. Localization

For the localization task, testing should ideally be performed using a ground truth segmented mask. However, as described earlier in the paper, we only have centroids of lesions to test against and as such we are using 225 known lesions and 225 randomly sampled non-lesion ROIs from the test scans. Convergence was quickly observed and the

Table 4: Architecture of 3D vanilla CNN for the classification problem.

Model Summary for vanilla CNN			
Layer	*Output*	*Kernel*	*Stride*
Input	(24,240,240,1)	NA	NA
Conv3D+BN	(24,120,120,64)	(3,3,3)	(1,2,2)
Conv3D+BN	(24,120,120,128)	(3,3,3)	(1,1,1)
MaxPool3D	(24,60,60,128)	(1,2,2)	(1,2,2)
Conv3D+BN(x2)	(24,60,60,256)	(3,3,3)	(1,1,1)
MaxPool3D	(24,30,30,256)	(1,2,2)	(1,2,2)
Conv3D+BN(x2)	(24,30,30,512)	(3,3,3)	(1,1,1)
MaxPool3D	(12,15,15,512)	(2,2,2)	(2,2,2)
Conv3D+BN(x2)	(12,15,15,1024)	(3,3,3)	(1,1,1)
MaxPool3D	(6,7,7,1024)	(2,2,2)	(2,2,2)
Conv3D+BN(x2)	(6,7,7,2056)	(3,3,3)	(1,1,1)
MaxPool3D	(3,3,3,2056)	(2,2,2)	(2,2,2)
GAP+Drop	(2056)	NA	NA
Dense	(64)	NA	NA
Concatenate	(65)	NA	NA
Output	(Batch size, 2)	NA	NA

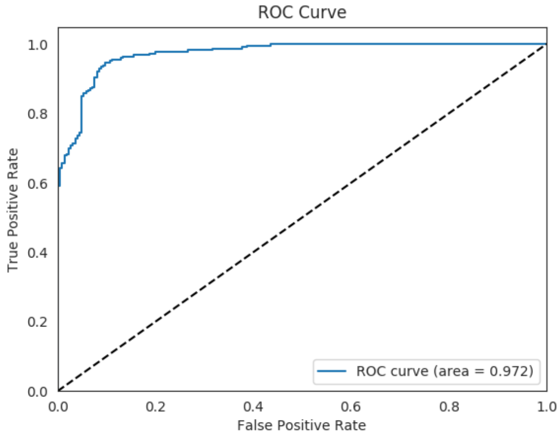

Figure 3: Localization problem: ROC curve for DenseNet.

Table 5: Localization problem: confusion matrix for test data (DenseNet).

Confusion Matrix			
	Prediction		
Actual	*No Lesion*	*Lesion*	*Class Accuracy*
No Lesion	203	22	90.2%
Lesion	13	214	94.3%

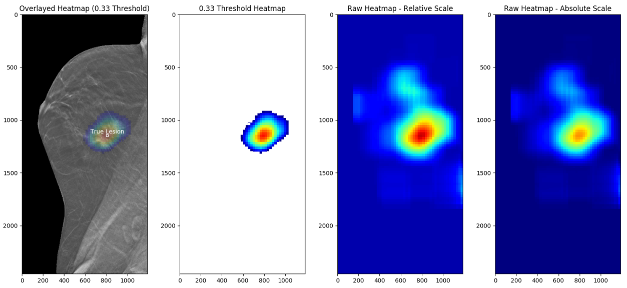
Overlayed Heatmap (0.33 Threshold)
True Lesion
0.33 Threshold Heatmap
Raw Heatmap - Relative Scale
Raw Heatmap - Absolute Scale
0
500
1000
1500
2000
0
200
400
600
800
1000

Figure 4: Post processing of an image in the training set. From left to right: Final heatmap overlaid on original image with true lesion marker, post-processed heat map with top 10% normalized, raw heat map normalized to image max and min, raw heat map absolute scale from 0 to 1.

best result on the validation set was achieved after only 10 epochs of training with an initial learning rate of 0.0001. We were able to achieve an AUC of 0.973 as shown in Figure 3 on the test set. Overall accuracy was 92.67% using a threshold of 0.33 determined from the ROC curve with a slight bias toward the true positive rate (TPR). The corresponding confusion matrix is shown in Table 5.

The localizer was then used on each of our sample images from Figure 1 to generate a heatmap. Since the raw heatmap consists of all likelihood scores from 0 to 1, post processing was performed to present only the predicted lesions in the image. Post processing included thresholding the raw map at 0.33 (from the ROC curve) and disregarding the rest. After thresholding, the heatmap is normalized to get the full spectrum of color on a relative scale to that image. Figure 4 shows the different steps in the post processing and the final heatmap.

Figure 5 shows some example heatmaps on the original images. These images show that the localizer is functional, but we do observe some false positives and false negatives.

7.1.1 Discussion and Future Work

Initial results demonstrate the feasibility of localization in DBT images. The sample heatmaps show that while our model is indeed capturing most of the lesions, we would prefer the heatmaps to be more sharply defined on the lesions. Aside from typical model optimization and tuning, this could be addressed through localization-specific data augmentation and using entire images to train the model. Each of these options are discussed further below.

In order to save time, lesion ROIs for the localizer were shared with the classification task. Although lesions for the localizer are indeed just the combination of both the benign and malignant lesions, the data augmentation techniques for the localization model could be chosen differently. Specifically, image shifting is an important augmentation technique for classification to allow for better generalization on

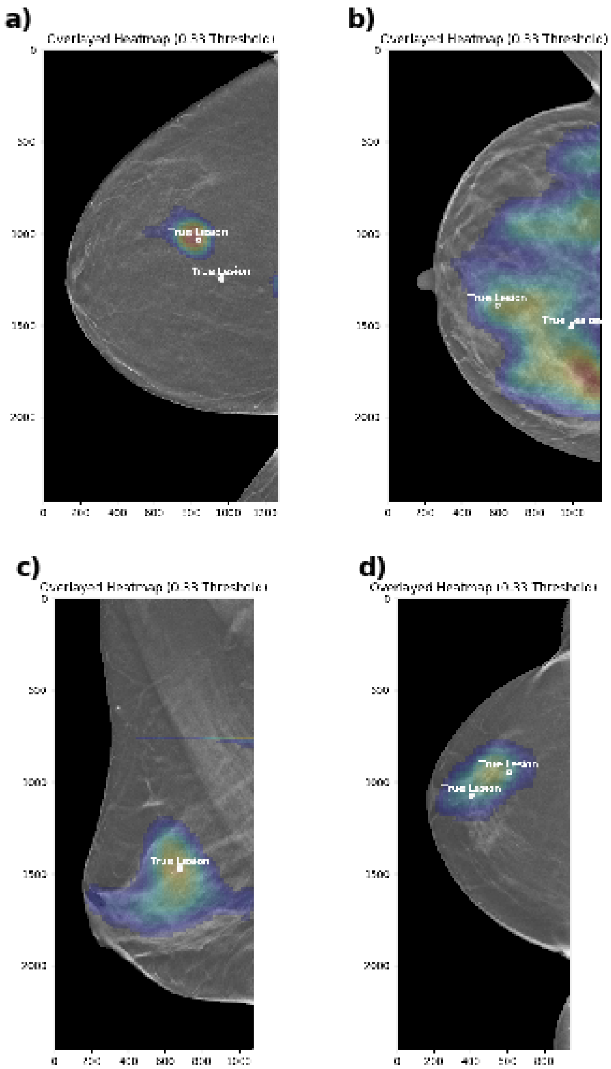
a)
Overlayed Heatmap (0.33 Threshold)
True Lesion
True Lesion
b)
Overlayed Heatmap (0.33 Threshold)
True Lesion
True Lesion
c)
Overlayed Heatmap (0.33 Threshold)
True Lesion
d)
Overlayed Heatmap (0.33 Threshold)
True Lesion
True Lesion

Figure 5: Heatmaps generated with post-processing, overlaid on original images, all from test set shown originally in Figure 1

non-centered ROIs. For localization, however, it might be better if the classifier was sensitive to lesion location within the ROI. By removing this augmentation step, or limiting shifts to the intended stride of the localization scanner, it may be possible to create a more sensitive heatmap.

Due to time constraints, training on entire scans was impractical and we instead performed limited random sampling from each scan to obtain the ROIs without lesions. With more time and resources, we would prefer to sample lot more non-lesion ROIs from each scan, which could possibly reduce the false positive rate.

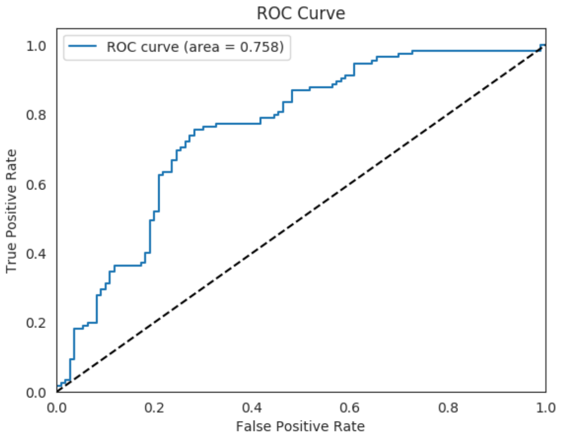

Figure 6: Initial results for the classification problem: ROC curve for vanilla CNN.

7.2. Classification

Although optimization and tuning of the classification stage are still in progress, we would like to report that initial results from the vanilla CNN, DenseNet and InceptionNet show AUC values of 0.758, 0.662 and 0.660 on the test set, respectively. Validation accuracy was monitored during training to save the best performing model prior to overfitting, which occurred at epoch 5, 9, and 10 for the vanilla CNN, DenseNet and InceptionNet, respectively. Table 6 shows the best performance of each model with the corresponding confusion matrix and accuracy on the test set. Figure 6 shows the ROC curves for the vanilla CNN. The ROC curve was used to select a threshold of 0.33, which resulted in the best observed accuracy of 73.3%.

Using the ROC curve to choose a threshold that achieves a sensitivity around 97-98%, our vanilla CNN is able to achieve a specificity of 30%, reducing the false positive rate by 13.6% compared to the BIRADS score only baseline, while maintaining a sensitivity of 97.4%. Table 7 shows the confusion matrix when the threshold is adjusted to 0.302 to achieve high sensitivity and Table 8 provides a comparison of the performance of our vanilla CNN to the BIRADS baseline.

7.2.1 Discussion and Future Work

Interestingly, the vanilla CNN produced significantly better results than DenseNet and InceptionNet, which could be due to the number of tunable parameters. It may be noted that the number of tunable parameters for the vanilla CNN is approximately 75 million, compared to 17 million and 2 million for DenseNet and Inception V2 respectively. With additional data, training time, and tuning, we are hopeful

Table 6: Top results from each model on testing data for the classification problem.

Confusion Matrix - vanilla CNN - 0.33 Threshold			
	Prediction		
Actual	*Negative*	*Positive*	*Class Accuracy*
Benign	77	33	70.0%
Malignant	27	88	76.5%

Confusion Matrix - InceptionNet - 0.42 Threshold			
	Prediction		
Actual	*Negative*	*Positive*	*Class Accuracy*
Benign	64	46	58.1%
Malignant	48	67	58.2%

Confusion Matrix - DenseNet - 0.43 Threshold			
	Prediction		
Actual	*Negative*	*Positive*	*Class Accuracy*
Benign	63	47	57.2%
Malignant	33	82	71.3%

Table 7: Classification problem: vanilla CNN with with high sensitivity threshold.

Confusion Matrix: vanilla CNN - 0.302 threshold			
	Prediction		
Actual	*Negative*	*Positive*	*Class Accuracy*
Benign	33	77	30.0%
Malignant	3	112	97.4%

Table 8: Classification problem: sensitivity and specificity comparison between our vanilla CNN and BIRADS baseline.

Baseline Comparison		
	Sensitivity	*Specificity*
Vanilla CNN	97.4%	30.0%
Baseline	97.8%	16.4%

that current results from all three models can be improved. However, we would like to emphasize that the initial results from the vanilla CNN are already better than those reported in existing literature using 3D CNNs. On a test set consisting of 110 benign and 115 malignant cases, our vanilla CNN achieved an AUC of 0.758 and accuracy of 73.3%. Previous work reported by Zhang et al[18] achieved an AUC of 0.6632 on a data set of 3018 benign and 272 malignant cases using 5-fold cross validation. Also, Wichakem et al[16] achieved 72.7% accuracy on a test set of 4 benign and 18 malignant cases. If the threshold is adjusted to maintain a sensitivity comparable to the BIRADS score only baseline, our vanilla CNN achieves a specificity of 30.0%, compared to 16.4% for the baseline. This is a significant improvement and alludes to the potential of our method to emerge as an useful aid to radiologists. However, 3D DBT machine learning models need to improve in order to take advantage of the spatial correlation between slices in order to improve classification performance, as has been achieved in other medical imaging such as CT for lung cancer. At this time, 2D CNNs operating on individual DBT slices have achieved the best results for classification; for example, an AUC of 0.91 has been reported by Samala et al[12].

In our work, we have not made any distinction between the MLO and CC views and treated all scans, whether MLO or CC, as unique samples, and one model was trained on data from both views. As an alternative, separate models could be trained for each view. The features extracted by each model could then be concatenated and fed to one final classifier, or alternately, we could train two completely independent models, each operating on an unique view, and the decisions from each model could be aggregated for the final output. This 2-channel feature extraction approach could be beneficial in case there are distinct features the models could learn from the MLO and CC views. Investigations along these lines will be conducted in future and results will be reported in a subsequent paper.

8. Conclusion

In this paper, we explored the feasibility of using 3D CNN models for localization and classification of lesions in DBT images. The localization model was able to output heatmaps with good accuracy and clarity, correctly identifying 93.0% of lesions while correctly identifying non lesions 91.1% of the time on our test set. The initial classification model was able to achieve an AUC of 0.758 with specificity and sensitivity of 70.0% and 76.5%, respectively. By adjusting the threshold to maintain false negative rates obtained with BIRADS, our vanilla 3D CNN was able to improve specificity on our dataset from 16.4% to 30% compared to using BIRADS alone. Given the limited amount of data and the need for an extremely high dimensional feature representation for proper classification, overfitting was a major issue for both tasks and special care was needed to address it.

The results of this paper show that CAD software can be developed for DBT images to aid in earlier detection of breast cancer with a reduced false positive rate. Compared to previous work using 3D CNNs on DBT images, this paper achieved significant results in both classification and localization of breast lesions but further improvement would be necessary for a practical CAD system. Potential future work has been outlined for both the localization and classification tasks to help achieve this goal.

References

[1] Carol E. DeSantis, Jiemin Ma, Mia M. Gaudet, Lisa A. Newman, Kimberly D. Miller, Ann Goding Sauer, Ahmedin Jemal, and Rebecca L. Siegel. Breast cancer statistics, 2019. *CA: A Cancer Journal for Clinicians*, page caac.21583, Oct. 2019. 1

[2] Raunak Dey, Zhongjie Lu, and Yi Hong. Diagnostic Classification Of Lung Nodules Using 3d Neural Networks. *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 774–778, Apr. 2018. arXiv: 1803.07192. 5

[3] Mahshid Ghoncheh, Zahra Pournamdar, and Hamid Salehiniya. Incidence and Mortality and Epidemiology of Breast Cancer in the World. *Asian Pacific Journal of Cancer Prevention*, 17(sup3):43–46, June 2016. 1

[4] grt123. Solution of the 'grt123' Team, 2017. 2

[5] Hugh Harvey, Edith Karpati, Galvin Khara, Dimitrios Korkinof, Annie Ng, Christopher Austin, Tobias Rijken, and Peter Kecskemethy. The Role of Deep Learning in Breast Screening. *Current Breast Cancer Reports*, 11(1):17–22, Mar. 2019. 2

[6] Kaggle. Data Science bowl 2017, 2017. 2

[7] Guixia Kang, Kui Liu, Beibei Hou, and Ningbo Zhang. 3d multi-view convolutional neural networks for lung nodule classification. *PLoS ONE*, 12(11), Nov. 2017. 5

[8] Eun-Kyung Kim, Hyo-Eun Kim, Kyunghwa Han, Bong Joo Kang, Yu-Mee Sohn, Ok Hee Woo, and Chan Wha Lee. Applying Data-driven Imaging Biomarker in Mammography for Breast Cancer Screening: Preliminary Study. *Scientific Reports*, 8(1):2762, Dec. 2018. 2

[9] Min Lin, Qiang Chen, and Shuicheng Yan. Network In Network. *arXiv:1312.4400 [cs]*, Mar. 2014. arXiv: 1312.4400. 5

[10] Bond M, Pavey T, Welch K, Cooper C, Garside R, and Et Al. Systematic review of the psychological consequences of false-positive screening mammograms. *Health Technology Assessment*, 17(13), Apr. 2013. 1

[11] Dragana Roganovic, Dragana Djilas, Sasa Vujnovic, Dag Pavic, and Dragan Stojanov. Breast MRI, digital mammography and breast tomosynthesis: Comparison of three methods for early detection of breast cancer. *Bosnian Journal of Basic Medical Sciences*, 15(4):64–68, Sept. 2015. 1

[12] Ravi K. Samala, Heang-Ping Chan, Lubomir Hadjiiski, Mark A. Helvie, Caleb D. Richter, and Kenny H. Cha. Breast Cancer Diagnosis in Digital Breast Tomosynthesis: Effects of Training Sample Size on Multi-Stage Transfer Learning Using Deep Neural Nets. *IEEE Transactions on Medical Imaging*, 38(3):686–696, Mar. 2019. 2, 8

[13] Philip Teare, Michael Fishman, Oshra Benzaquen, Eyal Toledano, and Eldad Elnekave. Malignancy Detection on Mammography Using Dual Deep Convolutional Neural Networks and Genetically Discovered False Color Input Enhancement. *Journal of Digital Imaging*, 30(4):499–505, Aug. 2017. 2

[14] Anna N. A. Tosteson, Dennis G. Fryback, Cristina S. Hammond, Lucy G. Hanna, Margaret R. Grove, Mary Brown, Qianfei Wang, Karen Lindfors, and Etta D. Pisano. Consequences of False-Positive Screening Mammograms. *JAMA Internal Medicine*, 174(6):954–961, June 2014. 1

[15] Anna Vlahiotis, Brian Griffin, A Thomas Stavros, and Jay Margolis. Analysis of utilization patterns and associated costs of the breast imaging and diagnostic procedures after screening mammography. *ClinicoEconomics and Outcomes Research: CEOR*, 10:157–167, Mar. 2018. 1

[16] Itsara Wichakam, Jatuporn Chayakulkheeree, and Peerapon Vateekul. Deep multi-label 3d ConvNet for breast cancer diagnosis in DBT with inversion augmentation. In Xudong Jiang and Jenq-Neng Hwang, editors, *Tenth International Conference on Digital Image Processing (ICDIP 2018)*, page 334, Shanghai, China, Aug. 2018. SPIE. 2, 8

[17] Adam Yala, Constance Lehman, Tal Schuster, Tally Portnoi, and Regina Barzilay. A Deep Learning Mammography-based Model for Improved Breast Cancer Risk Prediction. *Radiology*, 292(1):60–66, May 2019. 2

[18] Xiaofei Zhang, Yi Zhang, Erik Y. Han, Nathan Jacobs, Qiong Han, Xiaoqin Wang, and Jinze Liu. Classification of Whole Mammogram and Tomosynthesis Images Using Deep Convolutional Neural Networks. *IEEE Transactions on NanoBioscience*, 17(3):237–242, July 2018. 2, 8