Breast Cancer Classification Using Digital Breast Tomosynthesis Images

Emily Mui (em4449), Sanjay Subramanian (ss14383)

1. Introduction

Breast cancer is the most commonly diagnosed cancer worldwide, with approximately one in eight American women receiving a breast cancer diagnosis in their lifetime (Buda et al., 2021). Early diagnosis has been repeatedly shown to reduce disease burden and mortality, making it imperative to develop screening modalities which can identify abnormalities before clinical symptoms appear. The classical imaging technique employed for breast cancer screening is full-field digital mammography, in which two-dimensional X-ray views are imaged for interpretation by radiologists (Bai et al., 2021). While 2D mammography has led to a significant reduction in mortality as a result of regular screening, three-dimensional techniques have emerged as a new gold standard in screening capabilities, with increased sensitivity and reduced recall rates among some of the improvements offered (Vedantham et al., 2015). Traditional 2D views can lead to tissue artefacts from the summation of healthy dense tissue, which can mimic lesions or tumors on scans. Digital breast tomosynthesis (DBT), an emerging 3D technology, aims to solve this issue by using reconstructions of low-dose X-ray projection views acquired over an angular arc around the breast. These projection views can be used to build a z-stack of parallel slices in the same standard formats as those found in 2D mammography – bilateral craniocaudal (CC) and mediolateral oblique (MLO). Resolution and contrast are largely maintained while reducing tissue overlap. Less than one-third of Mammography Quality Standards Act (MQSA)-accredited breast imaging units in the US are DBT units, with cost, time, and demographic factors affecting widespread DBT adoption. Due to the 3D nature of DBT images and the number of slices and angles involved, it can take radiologists more than double or triple the amount of time to make assessments when compared to 2D mammography (Samala et al., 2018). This discord between potential and implementation presents an opportunity to develop best practices for DBT technologies using another re-emerging technology: deep learning.

The development of new deep and machine learning methods has been demonstrated to detect malignancies on traditional mammograms at the level of or even better than trained radiologists. Convolutional neural networks (CNNs) for classification have been used to reduce reading times without significantly affecting sensitivity, specificity, or recall rates, although these systems have mainly been used to support radiologists rather than being employed as independent diagnosis tools.

The lack of consensus and standardization surrounding DBT techniques and DBT-guided biopsies makes assessment of quality difficult when compared to established 2D methods. Most current classification studies using transfer learning approaches adapt weights based on 2D mammography images, followed by training on synthetic 2D reconstructions of 3D DBT images (Kim et al., 2016). Most studies of this format have reported AUC scores between 0.8 and 0.9 (Yousefi et al., 2018). In general, patient-matched DBT/mammography data is limited, making direct comparisons between the two modalities difficult. This makes it difficult to determine to what extent DBT images provide advantages over traditional mammograms.

In order to assess the quality of DBT images without the need for inclusion of mammography, this study leverages one of the largest public repositories of DBT images from The Cancer Imaging Archive (TCIA). While deep learning techniques can be extremely useful in image classification studies, lack of interpretability of features and restrictions of transfer learning from models pretrained on unrelated images can make training difficult.

2. Hypothesis

In certain cases, traditional machine learning models can outperform CNNs when feature sets are kept constant. Here we attempt to find out whether classical CNN models outperform traditional ML models trained on extracted features from these CNNs. Specifically, we extract features from each layer of the selected pretrained models and use them as features in a random forest model to 1. Determine which layers' features are most important for breast cancer classification and 2. Determine whether ensemble tree-based machine learning techniques, which are generally robust for classification tasks, can outperform CNNs.

3. Data

The dataset that was used in this project is publicly available on The Cancer Imaging Archive web site (Buda et al., 2021). It consists of 22,032 digital breast tomosynthesis (DBT) images from 5,060 participants that were collected from the Duke Health system between January 2014 and January 2018. Two radiologists with 18 and 25 years of experience assigned the patients' images with one of four labels: normal, actionable, benign, and cancer. It is important to note that the labels were assigned by patient and not by the individual image. For example, if a patient's left breast was designated to be actionable, both the left and right breast views would be classified as actionable. This could confound the model training, as it is unclear which side actually contains the suspicious area of interest, and so the images labeled as actionable were dropped.

The original dataset came with train/validation/test sets. However, the validation and test sets did not have labels. The train set was therefore used to generate 70/20/10 train/validation/test splits, with no patient overlap between the sets and the label stratified among the splits. Having no patient overlap between the data splits was especially important in this case because each normal patient has two views per breast for a total of four images, while each benign/cancer patient has two images for the biopsy. Example images of the CC and MLO views are shown in Figure 1. By ensuring that views of the same patient did not appear in multiple dataset splits, the potential for data leakage was minimized.

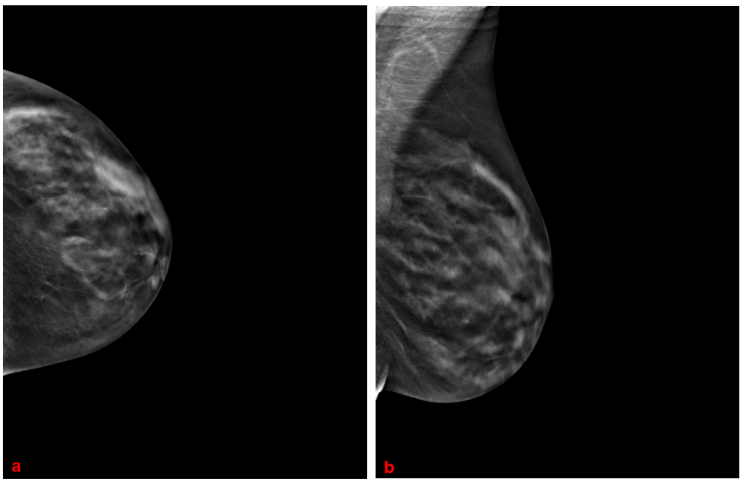

Figure 1: **(a)** Example CC view. **(b)** Example MLO view.

Table 1 shows the breakdown of the number of images included in each set. With such few positive labels in breast cancer screening, a binary classification task of normal vs. presence of tumor (benign + cancer) was performed. Even with combining the benign and cancer labels, the final datasets were still extremely unbalanced and so various techniques to alleviate the imbalance were employed and are expanded on in the Materials and Methods section.

	Train	Validation	Test
Benign/Cancer (1)	143	38	19
Normal (0)	12,760	3,648	1,824
Ratio	0.0112	0.0104	0.0104
Total Images	12,903	3,686	1,843

Table 1: Breakdown of the images in the dataset splits.

4. Materials and Methods

4.1 Data Preprocessing

The images were downloaded in DICOM format, a common image format for medical images that store both the image and patient data such as demographic information. With DBT images consisting of multiple views arranged into a z-stack, the first preprocessing step was to compress these 3D slices into synthetic 2D images that were then saved in JPG format. Images were then resized to 224 x 224 pixels and saved in this cropped format for more efficient model training and to stop jobs from being killed due to GPU inactivity. The images were normalized and augmented with various transforms from the torchvisions package, including RandomRotation, RandomResizedCrop, RandomErasing, RandomHorizontalFlip, and ColorJitter, to help with the extreme imbalance in the labels. Up/downsampling were also performed with WeightedRandomSampler from torch's sampler package, with different numbers of samples tested. Weighted binary cross-entropy was minimized in training to help address the class imbalance.

4.2 Models

For the pretrained model, a variety of models including pretrained ResNet-18, ResNet-34, and a simple CNN were evaluated for their performance on the validation set. For the two ResNet models, the number of input channels was changed to 1 to reflect the grayscale image input and the number of outputs of the final fully connected layer was updated to 2. The CNN architecture consisted of five convolution layers, with the first two layers having a stride of 2 and the following two layers having a stride of 3 and a pooling of 1. All convolution layers had a kernel size of 3 and used ReLU activation. After the convolution layers, there was an average pooling layer and a fully connected layer with two outputs for our binary classification problem. Hyperparameter tuning for these models included the learning rate, number of epochs, batch size, L2-regularization, weight decay, choice of scheduler, and number of samples (for WeightedRandomSampler). Some results from the hyperparameter tuning are shown in Table 2.

For the random forest model, features were extracted at each layer of the ResNet-34 model and pooled into a one-dimensional vector of features representing each image. Earlier layers in CNN models tend to extract higher-level features, while later levels tend to extract more granular specific features. The importance of each of these layers was assessed using random forest training, with hyperparameter tuning performed on max depth, number of trees, and minimum leaf and split sizes.

5. Results

The pretrained ResNet-18, ResNet-34, and the simple CNN models did not have promising results on the test set. They did not predict any positive labels at all, and so the AUC was 0.5. An analysis of the effect of training dataset size on AUC was also performed (with the same percentage of positive labels across the dataset sizes), but with the model already not performing well on the full training

set, the conclusions from that study are limited. The plot of AUC vs. dataset fraction is shown in Figure 2.

Weight_decay	Learning_rate	Class AUC
0.01	1E-5	0.53
	1E-6	0.47
0.1	1E-5	0.53
	1E-6	0.49
0	1E-5	0.48
	1E-6	0.47
1	1E-5	**0.54**
	1E-6	0.5
2	1E-5	0.53
a	1E-6	0.49

Out_channels	Learning_rate	Class AUC
128	1E-5	0.54
256	1E-4	0.50
256	1E-5	**0.62**
256	1E-6	0.46
256	1E-7	0.46
b 512	1E-5	0.53

Table 2: **(a)** Tuning weight_decay and learning_rate for ResNet-18. **(b)** Tuning number of out_channels and learning_rate for CNN.

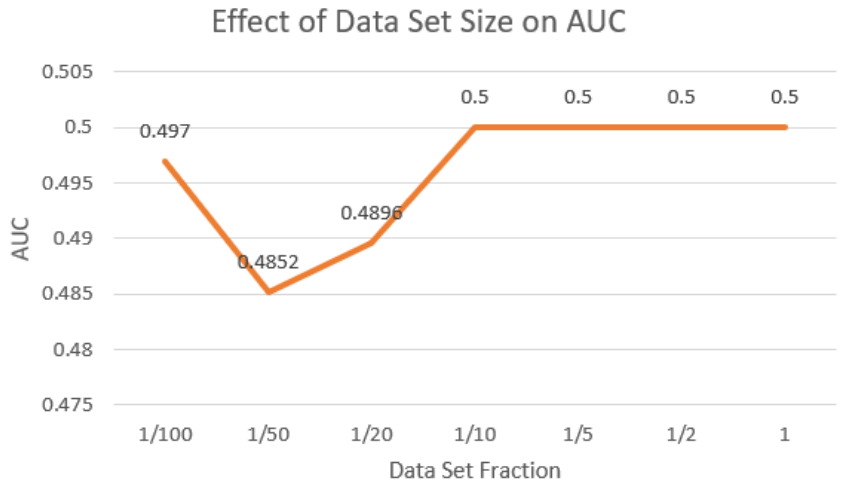

Figure 2: Effect of Different Training Dataset Sizes on AUC

For the feature extraction process, the output of each layer of interest was copied from the ResNet framework - this consisted of the four main layer blocks in addition to the average pooling layer. For the first four layers, as output was in the form of 3-dimensional arrays of channels and image vectors, an average pooling was used to condense the feature maps into one-dimensional arrays for input into the random forest. The pooling operation was modulated such that the final number of features was standardized to 512 from each layer. While feature maps from the first, second, and third layers tended to not be predictive, the fourth layer and average pooling layer of the ResNet models achieved AUCs as high as 0.75, which significantly outperformed the traditional ResNet models. These results can be seen in Table 3. Although not an extremely confident AUC, the 0.75 achieved

by the average pooling layer in the ResNet-18 model shows some promise in using feature extraction methods with traditional machine learning models as a more robust form of feature learning for image classification tasks.

Model	Layer	AUC
Resnet18	1	0.44
	2	0.52
	3	0.49
	4	0.64
	Avgpool	0.75
Resnet34	1	0.46
	2	0.48
	3	0.52
	4	0.59
	Avgpool	0.68

Table 3: AUC Scores for random forest model trained on features extracted from various layers of ResNet-18 and 34

6. Discussion

Although performance of the models was not high, we were still able to address our overall question of whether classical CNN models outperform traditional machine learning models that were trained on extracted features from CNNs in this specific area of using DBT images to classify tumor occurrence. Extracting features from various layers in a ResNet-34 and training a random forest on those extracted features had much better performance on our test set, resulting in an AUC of 0.75, compared to an AUC of 0.5 for our other models. Although this is a substantial improvement in AUC, it is important to note that the severe class imbalance seen in this dataset could affect the actual models and so in a more balanced dataset, a traditional CNN may perform better than a feature extraction method.

One major complication for this project was the severe class imbalance observed in the labels. Even with combining the benign and cancer labels to classify normal vs. presence of tumor, there were still only 200 positive images compared to 18,232 normal images, a ratio of 0.011. Techniques such as weighted cross-entropy, data augmentation, and upsampling were employed in the models to address the imbalance, but had minimal effect. Since the images are obtained through screenings for a cancer that has a prevalence of around 1-2% (Siegel et al., 2021), the positive labels would already be limited in traditional mammography images, let alone the newer DBT modality. Therefore, in order to obtain additional positive labels for DBT images, more institutions would need to adopt the use of DBT imaging so that deep learning models have more data and thus have more positive labels to train on.

Another issue that may have led to our models' poor performance on DBT images is that the DBT imaging technique produces slices that when viewed together, provide better insight into the underlying structure of the breast. However, when we compressed and converted the DBT slices into a single JPG image, valuable information was likely lost in the process. We did also try converting the DICOM images into TIFF format, which have slightly better image quality than do JPG images, but still obtained similar results.

6.1 Future Work

The two main issues that we faced in this project were the class imbalance and handling the 3D DBT images and so future work could focus on addressing these issues. For the class imbalance, pretraining a model on the more abundant traditional mammography images and finetuning the transferred weights on the DBT images would help provide the model with additional data and generate better weights. This method has already been shown to have improved results over ImageNet pretrained models (Samala et al., 2019), and so applying it to this dataset could also yield better results.

As for the compression of the 3D DBT slices to fit into our 2D models, 3D models could be used to maintain the extra dimensionality of the information contained in DBT slices. Although 3D models have been tried (Doganay et al., 2020), there have not been too many studies performed on this task, and so this is an active area of research. Because DBT images do take around twice as long as traditional mammography images for radiologists to analyze (Samala et al., 2016), it would be beneficial to develop robust and accurate 3D models that can serve as a guide for radiologists to make their diagnoses. With more widespread adoption of DBT imaging techniques and better deep learning models, there is great potential for obtaining more accurate breast cancer diagnoses.

7. Contributions

EM: data download, data exploration, data preprocessing, pretrained and simple CNN model; SS: literature review, data exploration, feature extraction random forest model

References

Jun Bai, Russell Posner, Tianyu Wang, Clifford Yang, and Sheida Nabavi. Applying deep learning in digital breast tomosynthesis for automatic breast cancer detection: A review. *Medical Image Analysis*, 71, 07 2021. doi: 10.1016/j.media.2021.102049.

Mateusz Buda, Ashirbani Saha, Ruth Walsh, Sujata Ghate, Nianyi Li, Albert Świecicki, Joseph Y. Lo, and Maciej A. Mazurowski. A data set and deep learning algorithm for the detection of masses and architectural distortions in digital breast tomosynthesis images. *JAMA Network Open*, 4(8): e2119100–e2119100, 08 2021. doi: 10.1001/jamanetworkopen.2021.19100.

Emine Doganay, Yahong Luo, Long Gao, Puchen Li, Wendie Berg, and Shandong Wu. Performance comparison of different loss functions for digital breast tomosynthesis classification using 3d deep learning model. *Medical Imaging*, 11314, 2020. doi: 10.1117/12.2551373.

Dae Hoe Kim, Seong Tae Kim, and Yong Man Ro. Latent feature representation with 3-d multi-view deep convolutional neural network for bilateral analysis in digital breast tomosynthesis. *IEEE*, 2016. doi: doi.org/10.1109/ICASSP.2016.7471811.

Ravi K. Samala, Heang-Ping Chan, Lubomir Hadjiiski, Mark A. Helvie, Jun Wei, and Kenny Cha. Mass detection in digital breast tomosynthesis: Deep convolutional neural network with transfer learning from mammography. *Medical Physics*, 43(12):6654–6666, 2016.

Ravi K Samala, Heang-Ping Chan, Lubomir M Hadjiiski, Mark A Helvie, Caleb Richter, and Kenny Cha. Evolutionary pruning of transfer learned deep convolutional neural network for breast cancer diagnosis in digital breast tomosynthesis. *Physics in Medine and Biology*, 63(9), 04 2018. doi: 10.1088/1361-6560/aabb5b.

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, 2019. doi: 10.1109/TMI.2018.2870343.

Rebecca L. Siegel, Kimberly D. Miller, Hannah E. Fuchs, and Ahmedin Jemal. Cancer statistics, 2021. *CA: A Cancer Journal for Clinicians*, 71(1):7–33, 01 2021. doi: 10.3322/caac.21654.

Srinivasan Vedantham, Andrew Karellas, Gopal R. Vijayaraghavan, and Daniel B. Kopans. Digital breast tomosynthesis: State of the art. *Radiology*, 277(3), 11 2015. doi: 10.1148/radiol.2015141303.

Mina Yousefi, Adam Krzyżak, and Ching Y. Suen. Mass detection in digital breast tomosynthesis data using convolutional neural networks and multiple instance learning. *Computers in Biology and Medicine*, 96, 05 2018. doi: 10.1016/j.compbiomed.2018.04.004.