


2023

Breast density classification using deep learning

Conrad Thomas Testagrose
ctestagrose95@gmail.com

Follow this and additional works at: <https://digitalcommons.unf.edu/etd>

 Part of the [Artificial Intelligence and Robotics Commons](#), [Data Science Commons](#), and the [Other Biomedical Engineering and Bioengineering Commons](#)

Suggested Citation

Testagrose, Conrad Thomas, "Breast density classification using deep learning" (2023). *UNF Graduate Theses and Dissertations*. 1178.

<https://digitalcommons.unf.edu/etd/1178>

This Master's Thesis is brought to you for free and open access by the Student Scholarship at UNF Digital Commons. It has been accepted for inclusion in UNF Graduate Theses and Dissertations by an authorized administrator of UNF Digital Commons. For more information, please contact [Digital Projects](#).

© 2023 All Rights Reserved

BREAST DENSITY CLASSIFICATION USING DEEP LEARNING

by

Conrad Testagrose

A thesis submitted to the
School of Computing
in partial fulfillment of the requirements for the degree of

Master of Science in Computer and Information Sciences

UNIVERSITY OF NORTH FLORIDA
SCHOOL OF COMPUTING

April, 2023

Copyright (©) 2023 by Conrad Testagrose

All rights reserved. Reproduction in whole or in part in any form requires the prior written permission of Conrad Testagrose or designated representative.

The thesis “Breast Density Classification Using Deep Learning” submitted by Conrad Testagrose in partial fulfillment of the requirements for the degree of Master of Science in Computing and Information Sciences has been

Approved by the thesis committee:

Date:

Dr. Xudong Liu
Thesis Advisor and Committee Chairperson

Dr. Indika Kahanda
Committee Member

Dr. Robert Maxwell
Committee Member

ACKNOWLEDGEMENTS

I would like to thank both the faculty from the University of North Florida and the Mayo Clinic for the tremendous amount of support that was provided to me throughout my time working on this thesis. I would specifically like to thank my advisor Dr. Xudong Liu, Dr. Indika Kahanda, Dr. Robert Maxwell, Dr. Richard White, Dr. Barbaros Erdal, Dr. Multu Demirer, and Dr. Vikash Gupta for all the support provided and knowledge gained. I would also like to thank Dr William Klostermeyer and Dr. Sherif Elfayoumy for their guidance and input on all the work that was carried out while working on my thesis at the Mayo Clinic.

CONTENTS

List of Figures.....	vii
List of Tables.....	xi
Abstract	xii
Chapter 1 Introduction	1
Chapter 2 Background and Related Work.....	5
2.1 Mammography.....	5
2.1.1 Imaging Process and Techniques.....	7
2.2 Transfer Learning.....	7
2.3 Data.....	8
2.4 Deep Learning Algorithms.....	8
2.4.1 Convolutional Neural Networks.....	8
2.4.1.1 Inception V3	9
2.4.2 Vision Transformer.....	9
2.4.3 DeepLabV3.....	11
2.5 AdaBoost.....	11
Chapter 3 Breast Density Classification Using Inception V3 and Vision Transformer.....	13
3.1 Training Exclusively on Craniocaudal Mammograms.....	13
3.1.1 Inception V3 CC/MLO vs Inception V3 CC Only.....	14
3.1.2 Inception V3 CC Only vs ViT CC Only.....	15
3.2 LCC and RCC image Concatenation.....	16
3.2.1 Inception V3 CC Only vs. Inception V3 CC Concatenated ..	18
3.2.2 ViT CC Only vs. ViT CC Concatenated.....	19

3.2.3	Inception V3 Concatenated vs. ViT CC Concatenated	21
3.2.4	AdaBoosting Inception-V3 for Comparison to ViT	22
3.2.5	Use of AdaBoost to Identify Noisy Labels	23
Chapter 4	Breast Density Segmentation Using DeepLabV3	26
4.1	Imaging Data	26
4.2	Segmentation Development Set	28
4.3	Verified Trial Set	30
4.4	Calculation of Density Ratio from Segmentation Masks	30
4.5	Threshold Determination	31
4.6	Probability Distribution Methodology	32
4.7	Results	32
Chapter 5	Conclusions and Future Work	37
5.1	Future Work	38
References	38
Appendix A	Related Publications, Achievements, and Deliverables	45
Appendix B	Additional Results	46

FIGURES

1.1	The four breast density categories outlined by BI-RADS in order of increasing breast density.	1
2.1	Each of a patient’s breasts are imaged from both the Mediolateral Oblique (MLO) and Craniocaudal (CC) orientations.	5
2.2	Visual example of the different mammogram image types. . .	6
2.3	General overview of the transformer architecture	10
3.1	AUC/ROC curve for Inception V3 trained on CC image orientation only and the AUC/ROC curve for Inception V3 trained on both CC and MLO image orientations	14
3.2	AUC/ROC curve for Inception V3 trained on CC image orientation only and the AUC/ROC curve for Inception V3 trained on both CC and MLO image orientations	16
3.3	Resulting image after concatenation of a patient’s LCC and RCC mammogram images	17
3.4	AUC/ROC curve for Inception V3 trained on CC image orientation only and the AUC/ROC curve for Inception V3 trained on concatenated CC images	18
3.5	AUC/ROC curve for ViT trained on CC image orientation only and the AUC/ROC curves for ViT trained on concatenated CC images at size 299 x 299 and 768 x 768	20
3.6	AUC/ROC curve for Inception V3 trained on concatenated CC images and the AUC/ROC curve for ViT trained on concatenated CC images	21

3.7	AUC/ROC curves for Inception V3, ViT, and the AdaBoosted LIC Classifier	23
4.1	MeVisLab GUI as it was presented to the radiologist for segmentation mask development. The GUI utilized a threshold based slider to highlight the pixels to be included in the mask. These pixels were highlighted in red.	27
4.2	The use of a threshold-based slider has the tendency to include undesired regions of the mammogram in the mask. Regions of the mammogram that were undesired such as the radiographic labels and regions of high exposure from the curvature of the breast were manually removed.	27
4.3	Pipeline for the development and evaluation of the breast-density segmentation algorithm. Segmentation training, validation, and test sets were created using 688 images from 329 patients from a total data set of 37,284 images from 17,625 patients. The density metric thresholds were extracted from the Segmentation Development Set. After training a segmentation model, the algorithm was applied to a verified set of 3,205 images from 1,522 patients. Applying the extracted thresholds to the density metric output by the model on the verified trial set, I determined the accuracy, a probability distribution, and a population distribution. The resulting data can then be used to display linear and probability scales to the radiologist using the automated tool.	29

4.4	Original patient RCC mammogram image (left) with its corresponding segmentation map overlaid (right). Segmented dense tissue (yellow) and the segmented breast tissue (blue green) can be used to calculate the patients breast density for each of their breasts. Using this value from both breasts, the patient breast density can be determined. Segmentation maps can be displayed to the radiologist along with their placement on the linear scale and probability charts.	31
4.5	Patient based breast density distributions. Black dashed lines show the thresholds extracted from the segmentation development set. Clearer separation is exhibited between classes B and C. Class A and B exhibit more overlap in their distributions. This observation is maintained with classes C and D. Due to legislative guidelines, the distribution lends some insight into radiologist adjudication around BI-RADS class thresholds. Radiologists may be more confident or careful around the B/C thresholds than around the A/B and C/D thresholds.	34
4.6	BI-RADS class probabilities based on the patient breast density percentages calculated from segmentation maps. The probability curves are determined using a polynomial regression model developed using the patient density distributions. Using this probability curve, we can output class probabilities for a given patient density to the radiologist.	35

4.7	Sample display of the functionality of the semantic breast density segmentation algorithm. Segmentation maps can be overlaid over the original images and both quantitative metrics for each breast as well as a BI-RADS class assignment and class probabilities can be displayed to the user.	35
-----	---	----

TABLES

3.1	The results for accuracy metrics of Inception V3 trained on CC and MLO image orientations vs. Inception V3 trained on only the CC image orientation.	14
3.2	The results for accuracy metrics of Inception V3 trained on CC images vs. ViT trained on only CC images	15
3.3	The results for accuracy metrics of Inception V3 trained on CC images vs. Inception V3 trained on concatenated images .	18
3.4	The results for accuracy metrics of ViT trained on CC images vs. ViT trained on concatenated images of size 299 x 299 and 768 x 768	19
3.5	The results for accuracy metrics of Inception V3 trained on concatenated images vs. ViT trained on concatenated images.	21
3.6	The results for Inception V3 vs. AdaBoosting of the LIC classifier	22
3.7	Class transition count and class transition percentage after radiologist re-evaluation of 100 "hard" images	24
3.8	The results for accuracy metrics of Inception V3 trained on concatenated images vs. ViT trained on concatenated images.	25
4.1	Dice Coefficients Across 5-Folds of Segmentation Set	33
4.2	BI-RADS Classification Accuracy Using Thresholds	33
4.3	Density Thresholds	34

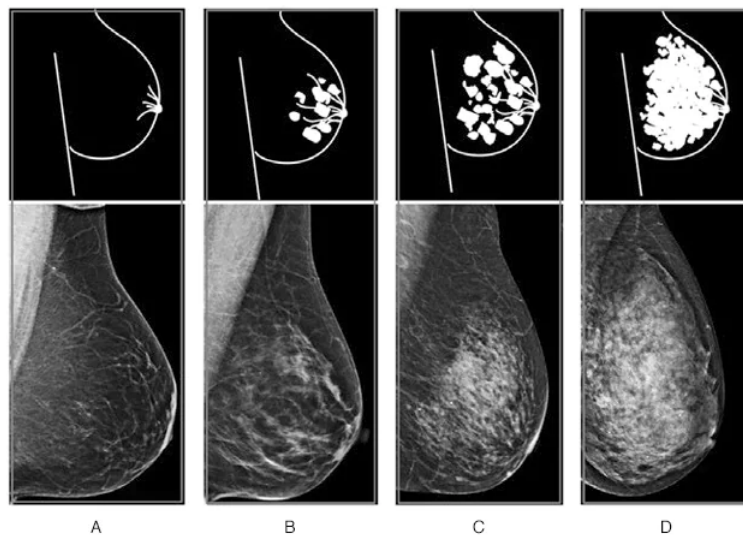
ABSTRACT

Breast density screenings are an accepted means to determine a patient's predisposed risk of breast cancer development. Although the direct correlation is not fully understood, breast cancer risk increases with higher levels of mammographic breast density. Radiologists visually assess a patient's breast density using mammogram images and assign a density score based on four breast density categories outlined by the Breast Imaging and Reporting Data Systems (BI-RADS). There have been efforts to develop automated tools that assist radiologists with increasing workloads and to help reduce the intra- and inter-rater variability between radiologists. In this thesis, I explored two deep-learning-based approaches on breast density classification. First, I developed and experimented with algorithms using deep learning models (such as Inception V3 and ViT) to classify patients according to BI-RADS using various types of digital mammograms. Second, with the need to provide not only such classification but also a quantitative measure of breast density to help standardize assessments across radiologists, I applied a deep learning based semantic segmentation model, DeepLabV3, to predict density percentages which then were used to provide a linear and probability scale.

CHAPTER 1

INTRODUCTION

Breast cancer is the most common form of cancer in women globally[38, 34]. Similar to other forms of cancer, earlier diagnosis and treatment can result in improved clinical outcomes for the patient. Breast density screenings are among the tools used to determine a patient's predisposed risk for breast cancer development. In a breast density screening, Radiologists visually assess patient mammogram images and visually assess the ratio of fibroglandular soft tissue to fatty tissue within the patient's breast. Based on their assessment, the Radiologist will then assign the patient to one of the four breast density categories outlined by the Breast Imaging Reporting and Data Systems (BI-RADS) [16] (Figure 1.1).



© MAYO FOUNDATION FOR MEDICAL EDUCATION AND RESEARCH, ALL RIGHTS RESERVED.

Figure 1.1: The four breast density categories outlined by BI-RADS in order of increasing breast density.

BI-RADS has subdivided breast density into a four-category scale according to the increasing relative contributions of dense fibroglandular tissue to the overall

breast tissue[16]. Most states in the United States have enacted breast density notification guidelines[2]. Many of these states further combine BI-RADS classes A and B into a non-dense category and BI-RADS C and D into a dense category. These guidelines legally require that radiologists notify patients within dense categories of increased levels of breast density of their anticipated increase in the risk for breast cancer development. These requirements to notification procedures also impact the recommended care plan for patients in these elevated breast density categories.

While BI-RADS is the most widely adopted breast-density assessment guideline [12], it is categorized as a subjective pattern based method for assessing patient breast density. The 5th edition of BI-RADS[14] has emphasized the removal of the quantitative assessed breast density quartiles present in the 4th edition of BI-RADS[16]. With this removal, BI-RADS has placed increased emphasis on the subjective text descriptions of the assessed density. BI-RADS explains that the change comes from the increased need to allow radiologists to assess density based on masking effect alongside the amount of dense tissue present [14]. However, this increased emphasis on subjective assessments does provide some drawbacks. Research has shown that the use of these subjective assessments for breast density can increase the inter-reviewer variability among radiologists [27, 11, 29] . This variability can have a significant impact on the patient's determined breast density and the provided treatment plan. Patient's laying on the thresholds between BI-RADS classes can have varying assessments that are dependent on the radiologist reviewing their mammogram images. The American College of Radiology has issued a statement in regard to this variability in subjective breast density assessments among providers, claiming that the use of subjective pattern-based breast density assessment guidelines can result in a significant reduction in the reproducibility of assessment results[1].

To help standardize assessment results alongside the increasing world population and the increasing breast cancer incidence rates [34], there has been a growing interest in the development of automated tools to assist radiologists with breast density screenings. The use of deep learning approaches to assist radiologists with the classification of patients into one of the four BI-RADS classes has seen growth in recent years [6]. Convolutional Neural Networks (CNN) are the most common deep learning architecture used to accomplish this classification task [39]. CNNs have been applied to both image classification and object detection tasks and have experienced widespread utilization since the 1990s[23]. The use of CNNs as a deep learning breast density classifier has the opportunity to assist radiologists with increasing workloads as well as to help standardize the results between different radiologists.

While CNNs are extremely useful for breast density classification, I argue that the assessment of breast density is not solely a classification task and relies on the subconscious segmentation of dense breast tissue from fatty breast tissue. The use of deep neural networks for breast density classification acts as a black box and the classifier uses any number of features to make its classification. This makes it difficult to extract the exact metric or feature used by the algorithm to make the assessment. Previous research has investigated the use of segmentation algorithms as a means to extract a quantitative density metric from mammogram images and utilize this metric to more objectively assist radiologists in breast density assessment. One such method is Cumulus[9], a semi-automated quantitative breast density assessment tool developed by researchers at the University of Toronto. This method utilizes a threshold set by a trained observer at the time of assessment to "segment" the dense pixels from the non-dense pixels and output a quantitative metric. While Cumulus provides radiologists with the means to reduce inter-reviewer variability between breast density assessments, its semi-automated

nature that requires trained users makes it much more challenging to work into radiological workflows. There have been efforts to develop fully automated quantitative breast density tools that aim to reduce the amount of human interaction necessary in outputting a density metric. Some fully automated tools, such as the tool developed by researchers at the Karolinska Institute, utilize ImageJ to output density as a continuous percentage [24] while other fully automated methods (e.g., LIBRA) utilize a fuzzy c-means based approach to provide a continuous breast density percentage[21]. Machine learning has also been used by STRATUS to provide comparative performance to the semi-automated Cumulus and provides both a continuous breast density percentage along with a corresponding BI-RADS class based on predefined cut-off points[17]. The few commercially available quantitative area-based breast density tools such as iCAD’s PowerLook[4] and Densitas[3] possess limited publicly available information on the actual methodologies used in their development.

In this thesis I aim to explore 1. the development and comparison of an Inception-V3[36] based breast density classifier with and without AdaBoost to a Vision Transformer (ViT)[15] based breast density classifier using a novel image pre-processing technique; and 2. The use of the DeepLabV3[10] semantic segmentation architecture for the development of a quantitative breast density assessment tool.

CHAPTER 2
BACKGROUND AND RELATED WORK

2.1 Mammography

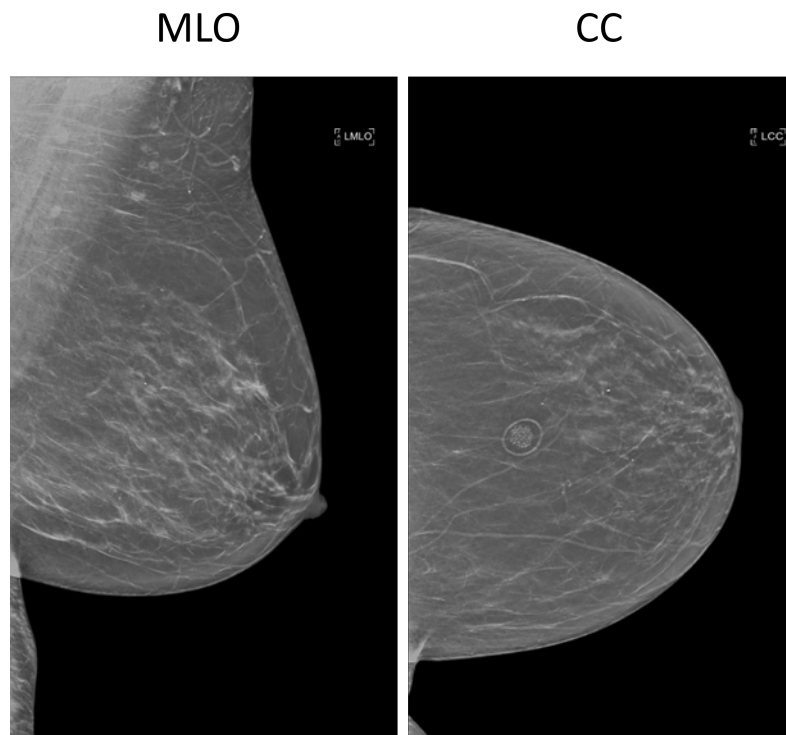


Figure 2.1: Each of a patient’s breasts are imaged from both the Mediolateral Oblique (MLO) and Craniocaudal (CC) orientations.

This thesis utilized mammogram images to develop deep learning breast density classification and segmentation algorithms. Mammography is an x-ray imaging technique used in the screening and detection of breast cancer and other diseases associated with the breast. The images produced by a mammogram show the internal tissues of the breast with darker shades of grey indicating lower density tissues and white or whiter shades of grey indicating higher density tissues. Mam-

mammograms are typically taken from the Craniocaudal (CC) and the Mediolateral Oblique (MLO) orientations (Figure 2.1). Radiologists will use CC and MLO mammograms from both a patient’s left and right breast to make a breast density assessment.

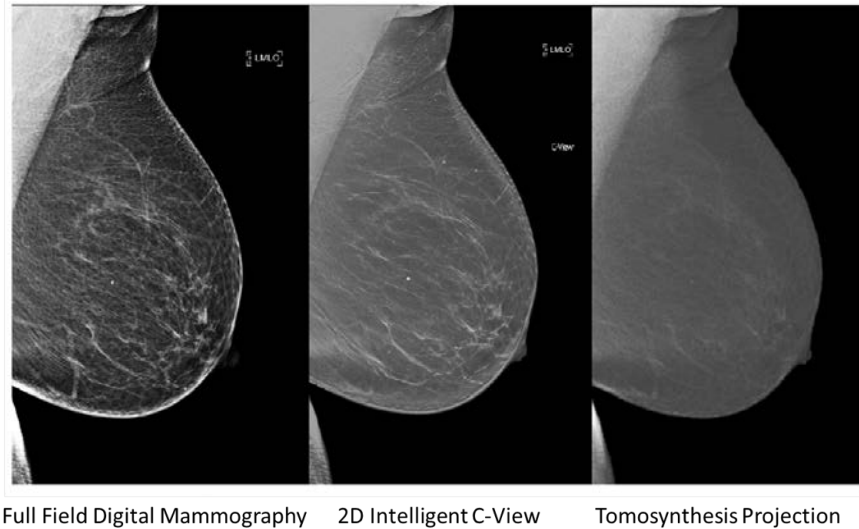


Figure 2.2: Visual example of the different mammogram image types.

The Mayo Clinic mammogram data set used for this thesis consisted of mammogram images from both the CC and MLO orientations for the right and left breast. The CC orientation is taken from a top-down perspective of the breast while the MLO orientation is taken from a side-to-side perspective. This thesis work has placed emphasis on the use of the CC mammogram images for the development of both classification and segmentation algorithms. The pectoral muscle present in the upper left or right quadrant of MLO mammogram images possesses a similar opacity to the dense tissue that is of interest [30, 28, 26]. By excluding MLO images, the avoidance of error introduced by pectoral muscle removal algorithms can be accomplished. MLO images also often fail to include the deeper tissues of the medial portion of the breast [8], resulting in more breast tissue being present in CC mammogram images.

2.1.1 Imaging Process and Techniques

During the imaging process, a patient’s breasts are compressed and an x-ray machine is then used to send x-rays through the breast to a detector on the opposite side. This detector converts the x-rays to signals that a computer uses to develop an image. The resulting 2D image is known as a Full Field Digital Mammogram. The development of newer technologies has provided patients with both C-View (Hologic, Marlborough, MA, <https://www.hologic.com/>) mammogram images and Tomosynthesis Projection images (Figure 2.2). Both C-View and Tomosynthesis Projection formats are becoming increasingly popular in routine mammographic examinations due to the reduction of both the breast compression time and the administered radiation dose[5].

2.2 Transfer Learning

All models used in this thesis utilized a technique known as transfer learning when available. In transfer learning, the weights of a publicly available model pre-trained on an exceptionally large and general data set like ImageNet[22], Microsoft COCO[25], or ILSVC[32] to a model used on a smaller and more specific data set. Transfer learning is particularly useful in data-constrained situations with smaller data sets as it helps to train the neural network on limited data, while often achieving higher accuracy with less training time[37]. The models of Inception-V3[36] and ViT[15] were pre-trained on the ImageNet[22] data set while the DeepLabV3[10] models used were pre-trained on a subset of the Microsoft COCO[25] data set.

2.3 Data

All the data sets used in this thesis were created using Mayo Clinic data. A total of 118,459 mammogram images from 9899 patients were sampled across multiple Mayo Clinic enterprise locations. The data was split based on the patient when creating the development, validation, and testing data sets. This was done to ensure that patients that were trained or validated on would not have samples present in the testing set.

2.4 Deep Learning Algorithms

For this thesis, the focus was primarily on the deep learning approaches used to develop breast density classification and segmentation algorithms. For classification, two different types of deep learning architectures were used: Convolutional Neural Networks (CNNs) and Vision Transformers (ViT). For segmentation, the DeepLabV3[10] semantic segmentation architecture was used.

2.4.1 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are deep learning algorithms that focus on images. They are feed-forward neural networks that consist of input, hidden, and output layers. CNNs are unique in that some of their hidden layers perform convolutions. Convolutional layers convolve over their receptive field of the input image passing the results to the next layer. The layers produce activation or feature maps of the input image. CNNs also possess pooling layers which aim to reduce the dimensions of the data output from multiple neurons of one layer into one neuron in the following layer. The two forms of pooling are known as max pooling and average pooling. Max pooling takes the max value of the outputs in the feature map while average pooling takes the average value. CNNs are often the go to deep

learning algorithm for image classification.

2.4.1.1 Inception V3

Inception-v3, a CNN developed by Google, is an updated version of GoogLeNet which was introduced in 2014 [35, 36]. The initial Inception model aims to avoid overfitting data from deep layers of convolutions by using an inception module. Multiple filters of differing sizes are used within an inception module. Each module consists of 1x1, 3x3, and 5x5 convolutions along with 3x3 max pooling. Inception-v1 introduced 1x1 convolutions before the 3x3 and 5x5 convolutions and a 1x1 convolutions after the 3x3 max pooling within the inception module. This was done to reduce the computationally expensive nature of the 5x5 convolution. These inception modules have undergone optimization over time resulting in higher efficiency, a deeper network without compromising speed, and the use of auxiliary classifiers to regularize. The use of inception modules also promotes the extraction of features of varying scales due to the differing convolutional filter sizes. Due to this, Inception-v3 could be useful in the extraction of the dense regions of the breast while also identifying the whole breast and providing better performance regarding accuracy metrics.

2.4.2 Vision Transformer

Transformers have seen successful growth and usage in the area of Natural Language Processing (NLP). Transformers utilize a concept known as attention which mimics cognitive attention. Some parts of input data are enhanced while others are diminished allowing for the algorithm to learn the part of the data that is more important than the others. Only recently has the concept of using Transformers on 2D image data been proposed in the form of Vision Transformers (ViTs) [15]. ViTs work by first splitting an input image into patches of equal size. These patches

are then flattened and positional embeddings are added to the linear embeddings of these flattened patches. The Transformer Encoder is then fed the sequence of patches. Following the Transformer Encoder, a Multi-Layer Perceptron head is used to classify the resulting output from the Transformer Encoder (Figure 2.3). While ViTs require an immense amount of data, they have been shown to outperform traditional CNNs on datasets such as ImageNet. Therefore, due to the characteristics of the ViT framework, the use of ViT for breast density classification may provide and increase in performance over a CNN.

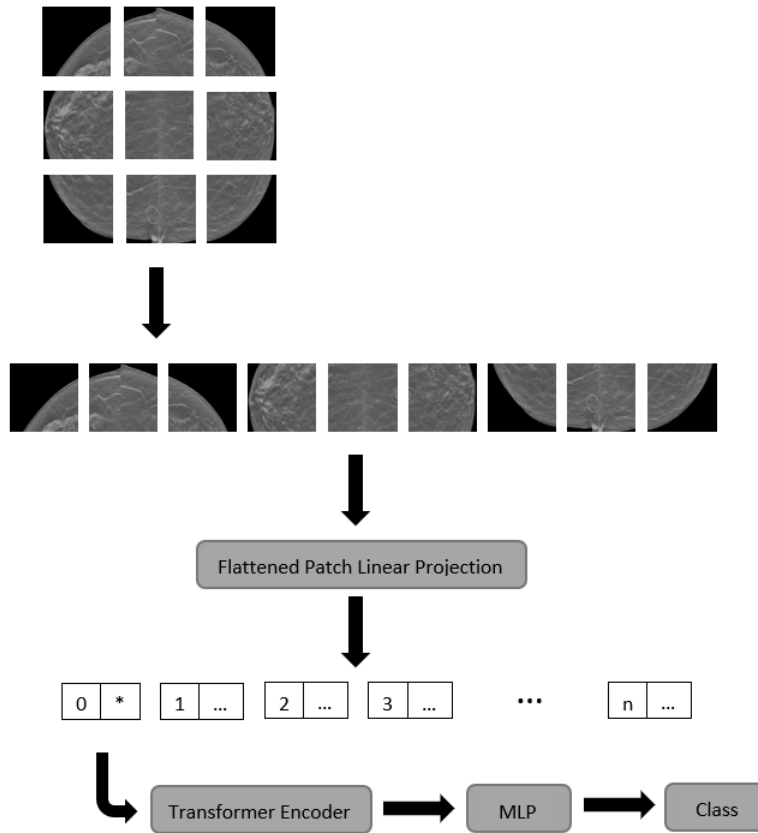


Figure 2.3: General overview of the transformer architecture

2.4.3 DeepLabV3

DeepLabV3 is an image semantic segmentation algorithm developed by Google [10]. DeepLabV3 works by extracting features from the backbone convolutional neural network. The size of the feature map is controlled using atrous convolutions in the last few layers of the backbone network to increase the dilation rate at each layer. Atrous Spatial Pyramid Pooling (ASPP) is then used to classify each pixel as corresponding to an individual class. A 1x1 convolution is applied to the final output from the ASPP network resulting in the final segmented mask from the algorithm. The filter's field of view becomes larger which in turn allows for better semantic segmentation. The characteristics of the atrous convolutions could provide a higher quality segmentation of the dense tissue of the breast. Due to its use of atrous convolutions and ASPP along with pretrained public availability through PyTorch, DeepLabV3 was selected over similar algorithms such as FCN[33], UNet[31], and SegNet[7].

2.5 AdaBoost

AdaBoost is a popular boosting algorithm in machine learning. Most implementations of AdaBoost focus on traditional machine learning algorithms rather than deep learning algorithms. Pseudocode outlining the AdaBoost algorithm, as described in [18] is displayed below:

At the start of the algorithm all the samples in the training set are weighted equally. The classifier is fit with the training set and the weighted error is calculated. An update parameter known as α_j is calculated and used to update the weights of the samples in the training data. The weights of the misclassified samples are increased while the weights of correctly classified samples are decreased. process continues until an optimal error is achieved or a specified number of itera-

Algorithm 1 AdaBoost Algorithm

1. Initialize initial weights $w_i = 1/n, i = 1, 2, \dots, n$.
2. For $m = 1$ to M :
 - (a) Fit classifier $T^{(m)}(x)$ with training set using weights w_i .
 - (b) Compute:

$$err^m = \frac{\sum_{i=1}^n w_i II(c_i \neq T^m(x_i))}{(\sum_{i=1}^n w_i)}$$

- (c) Compute:

$$\alpha^{(m)} = \log \frac{1 - err^{(m)}}{err^{(m)}} + \log(K - 1)$$

- (d) Set $w_i \leftarrow w_i \cdot \exp[\alpha^{(m)} \cdot II(c_i \neq T^m(x_i))], i = 1, 2, \dots, n$.
 - (e) Re-normalize w_i
 3. Output $C(x) = \operatorname{argmax}[\sum_{m=1}^M \alpha^{(m)} \cdot II(T^{(m)}(x) = k)]$.
-

tions is reached. The trained classifiers form an ensemble, each of these classifiers are known as weak classifiers which have learned their respective weak hypotheses. This ensemble of classifiers is then used to perform classification with the notion that the ensemble of weak classifiers will outperform a single strong classifier without AdaBoost. It should also be noted that the accuracy of each classifier needs to be better than random guessing for $\alpha^{(m)}$ to be positive. Therefore, $(1 - err^{(m)} > \frac{1}{K})$ must hold true where K is the number of classes.

CHAPTER 3

BREAST DENSITY CLASSIFICATION USING INCEPTION V3 AND VISION TRANSFORMER

With ViT being a newer technology that boasts exceptional performance results [15], the comparison of a ViT-based breast density classification algorithm to an Inception-V3 based breast density classification algorithm was carried out. While performing this comparative study, the identification of a novel mammogram image pre-processing technique that concatenated a patient’s left and right breast image was identified. Use of this technique in this comparison provided patient-based results for more objective comparison of the algorithms. The exploration of AdaBoosting and inception-based algorithm was also performed to determine if this boosting technique provides any benefits over the use of a ViT.

3.1 Training Exclusively on Craniocaudal Mammograms

Prior to comparing Inception-V3 to ViT it was necessary to ensure that training on only the CC image orientation would provide comparative results to training on both CC and MLO images. Excluding the MLO image orientation from the set of mammogram images used to train reduces the size of the resulting data set. While excluding the MLO images would reduce the size of the data set used to develop the algorithm, it would aid in the reduction of introduced error from pectoral muscle removal algorithms. The use of the solely the CC images could also provide a more objective breast density assessment as MLO images lack the deep tissues of the medial portion of the breast that is included in the CC image orientation [8]

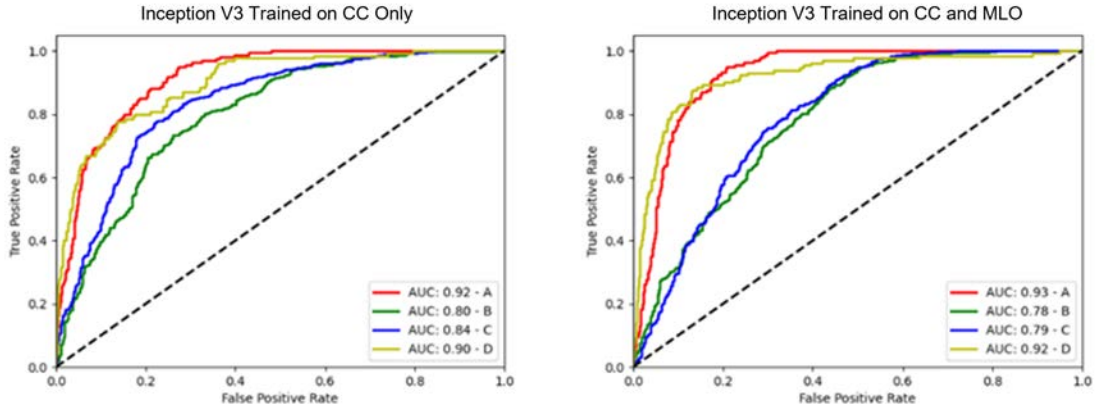


Figure 3.1: AUC/ROC curve for Inception V3 trained on CC image orientation only and the AUC/ROC curve for Inception V3 trained on both CC and MLO image orientations

3.1.1 Inception V3 CC/MLO vs Inception V3 CC Only

Using Inception V3 as a baseline, experiments were run to determine if training on exclusively CC images would achieve comparable results to Inception V3 trained on both CC and MLO images. The performance was compared using accuracy metrics (Table 3.1) and Area Under the Receiver Operator Characteristic Curve (AUC/ROC) (Figure 3.1). Both models were trained on their respective training sets and evaluated on the same test set which consisted of 4555 CC and MLO images.

Inception V3 CC/MLO vs Inception V3 CC Only				
Model	Validation Accuracy	Test Accuracy	Macro F1	Weighted F1
CC and MLO	0.73	0.72	0.70	0.72
CC Only	0.74	0.72	0.68	0.71

Table 3.1: The results for accuracy metrics of Inception V3 trained on CC and MLO image orientations vs. Inception V3 trained on only the CC image orientation.

Results from the experimentation show that training Inception V3 on exclusively the CC image orientation provides comparable results to training Inception V3 on both CC and MLO image orientations. Validation accuracy saw a 1 percent increase when using only C images while the test accuracy saw no change between using exclusively CC images or using both CC and MLO images. Macro F1 saw a 2 percent decrease and weighted F1 saw a 1 percent decrease when using exclusively CC images. Inception V3 trained exclusively on CC images showed a notable increase to the AUC for classes B and C, 2 and 5 percent respectively. Training with both CC and MLO images provided a 1 percent increase to AUC in classes A and D over training exclusively on CC images.

3.1.2 Inception V3 CC Only vs ViT CC Only

After ensuring that developing the classifier on solely CC images provides comparative performance to using both CC and MLO images, comparison of the performance of Inception-V3 and ViT using only CC images was performed. 5-fold cross validation was carried out for the following tests in order to compare both models more completely. Both models used an image size of 299 x 299 and ViT used a patch size of 32 x 32. The average accuracy, F1 scores, and AUC were recorded using the 5 folds of the data set.

Inception V3 CC Only vs ViT CC Only				
Model	Validation Accuracy	Test Accuracy	Macro F1	Weighted F1
Inception V3	0.74	0.73	0.70	0.73
ViT	0.75	0.74	0.71	0.74

Table 3.2: The results for accuracy metrics of Inception V3 trained on CC images vs. ViT trained on only CC images

The results for Vit against Inception V3 for training exclusively on the CC image orientation are shown in Table 3.2 and Figure 3.2. Using ViT provided on

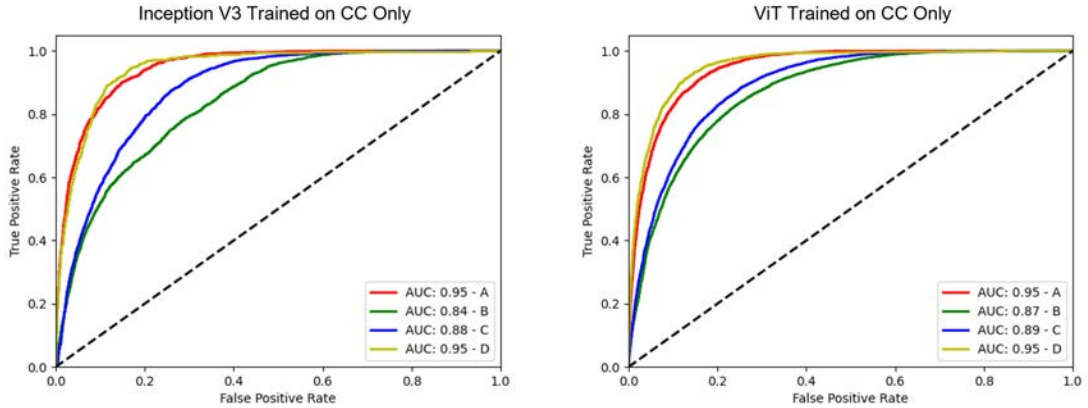


Figure 3.2: AUC/ROC curve for Inception V3 trained on CC image orientation only and the AUC/ROC curve for Inception V3 trained on both CC and MLO image orientations

average a 1 percent increase to all accuracy metrics tested. However, this increase was not determined to be significant. The AUC/ROC for ViT did show a significant result of a 3 percent increase for class B AUC ($p = 0.0499$) and a 1 percent increase to class C AUC ($p = 0.0372$) over Inception V3 on average. This result shows that ViT was significantly better at distinguishing between classes B and C.

3.2 LCC and RCC image Concatenation

Patients are assigned a breast density classification after a breast density screening by a radiologist. This classification serves as the ground truth label when training deep learning breast density classifiers. While carrying out the previous experiments, it was observed that there was a difference in the amount of dense tissue between a patient’s breasts. This discrepancy is due to natural asymmetry that is exhibited by biological systems such as the human body[19]. There are cases where one of a patient’s breasts can be classified as lower on the BI-RADS scale while the other breast could be classified as higher on the BI-RADS scale. When this situation occurs, the radiologist will assign a density classification to the patient

based on the breast with the higher density. This poses a challenge when training deep learning breast density classifiers on individual mammogram images. Due to the ground truth label being patient based, training on this label with individual images can introduce a degree of error into the development of these classifiers. For this reason we propose an image technique where we concatenate a patient’s left CC image (LCC) and right CC image (RCC) together. The below image displays a sample concatenated image.

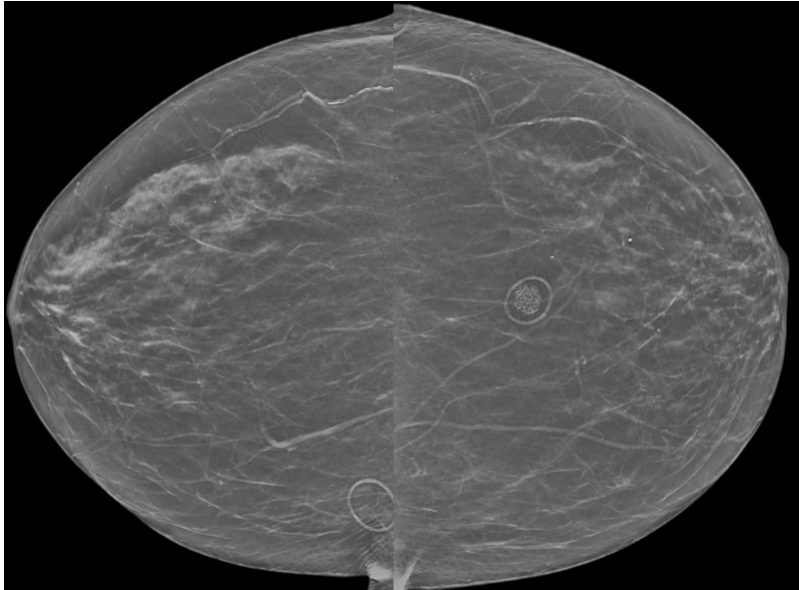


Figure 3.3: Resulting image after concatenation of a patient’s LCC and RCC mammogram images

The resulting image contains both breasts back-to-back in the center of view. To our knowledge, this image pre-processing technique has not been performed prior. We developed data sets of concatenated images and performed tests to determine if using this image pre-processing technique will increase the performance of our breast density classifiers. The data set of concatenated images consisted of 9899 images from 9899 patients. Tests were carried out using both Inception V3 and ViT.

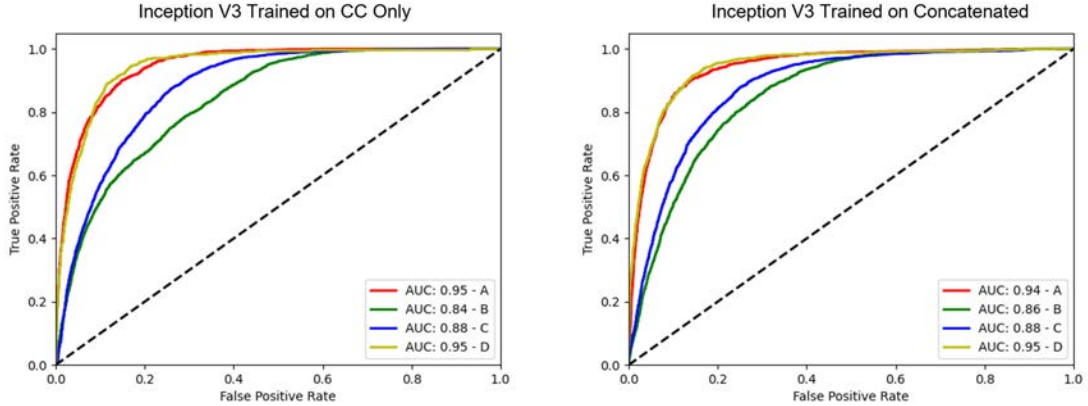


Figure 3.4: AUC/ROC curve for Inception V3 trained on CC image orientation only and the AUC/ROC curve for Inception V3 trained on concatenated CC images

3.2.1 Inception V3 CC Only vs. Inception V3 CC Concatenated

Initial tests using concatenated CC mammogram images were carried out using Inception V3. Using 5-fold cross-validation, models were trained on either exclusively images from the CC orientation or concatenated images. All images were resized to the 299 x 299 image size required by Inception V3.

Inception V3 CC Only vs Inception V3 Concatenated				
Model	Validation Accuracy	Test Accuracy	Macro F1	Weighted F1
Inception V3 CC Only	0.74	0.73	0.70	0.73
Inception V3 Concatenated	0.76	0.75	0.72	0.75

Table 3.3: The results for accuracy metrics of Inception V3 trained on CC images vs. Inception V3 trained on concatenated images

The results comparing Inception V3 using concatenated CC images and individual CC images are shown in table 3.3 and figure 3.4. Training an Inception V3 breast density classifier using concatenated CC images provided a significant increase of 2 percent to validation accuracy ($p = 0.0144$), test accuracy ($p = 0.0150$), macro F1 ($p = 0.0234$), and weighted F1 ($p = 0.0170$). The AUC/ROC for Inception V3 using concatenated images shows a decrease of 1 percent to class A

AUC and an increase of 2 percent to class B AUC. The changes to the AUC for Inception-V3 with concatenated images were not significant.

3.2.2 ViT CC Only vs. ViT CC Concatenated

Experiments using ViT and concatenated images were run following the experiments using Inception-V3. Tests utilized ViT models using a patch size of 32 x 32 and 5-fold cross-validation. Initial testing with a patch size of 32 x 32 appeared to provide more consistent results. All images were resized to 299 x 299 to ensure that a fair comparison of ViT to Inception-V3 could be conducted. We also ran tests of ViT with the image size increased to 768 x 768. One of the benefits of using a ViT is that we can more easily update the size of the image required by the network. It was hypothesized that increasing the image resolution closer to native resolution could provide an increase in algorithm performance.

ViT CC Only vs ViT Concatenated				
Model	Validation Accuracy	Test Accuracy	Macro F1	Weighted F1
ViT CC Only	0.75	0.74	0.71	0.74
ViT 299x299 Concatenated	0.76	0.76	0.73	0.76
ViT 768x768 Concatenated	0.76	0.76	0.72	0.76

Table 3.4: The results for accuracy metrics of ViT trained on CC images vs. ViT trained on concatenated images of size 299 x 299 and 768 x 768

The results comparing ViT using concatenated CC images and individual CC images are shown in table 3.4 and figure 3.5. Using a ViT with concatenated images provided on average a 1 percent increase to the validation accuracy but a 2 percent increase to the test accuracy, macro F1, and weighted F1. Of these increases to the performance metrics only the increase to test accuracy ($p = 0.0431$) and weighted F1 ($p = 0.0431$) were significant. When we increased the size of the concatenated images from 299x 299 to 768 x 768 and compare the performance between both

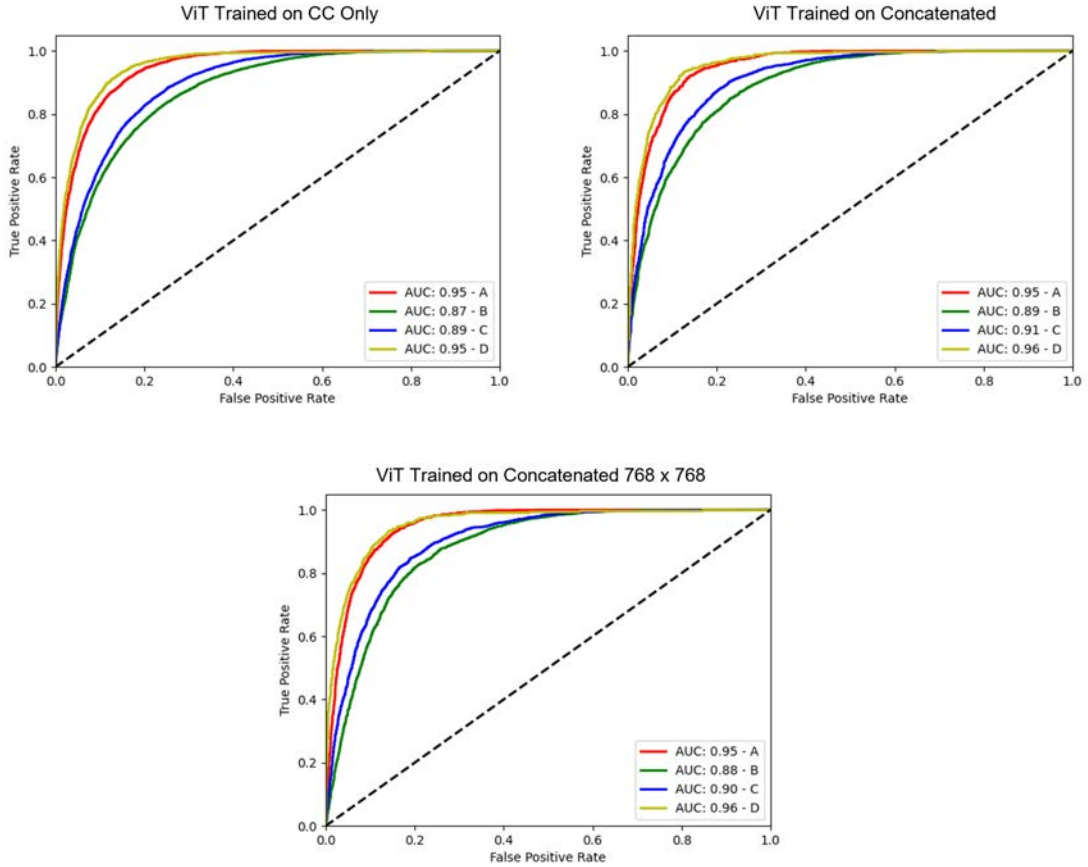


Figure 3.5: AUC/ROC curve for ViT trained on CC image orientation only and the AUC/ROC curves for ViT trained on concatenated CC images at size 299 x 299 and 768 x 768

ViTs, we see that there is no notable change in the performance outside of a 1 percent decrease to the macro F1 when using the larger image resolution. The AUC/ROC for the ViT trained on concatenated images of size 299 x 299 shows 2 percent increases to both class B and C AUC along with a 1 percent increase in class D AUC over individual CC images. Of these increases the change to class C ($p = 0.0064$) and class D (0.0161) AUC were the only significant changes. Increasing the concatenated image size to 768 x 768, the AUC for classes B and C were 1 percent lower than when using concatenated images at size 299 x 299.

3.2.3 Inception V3 Concatenated vs. ViT CC Concatenated

Following the aforementioned experimentation we then compared the use of concatenated images and Inception-V3 to ViT. Both networks used image sizes of 299 x 299 and 5 fold cross-validation for a fair comparison.

Inception V3 Concatenated vs ViT Concatenated				
Model	Validation Accuracy	Test Accuracy	Macro F1	Weighted F1
Inception V3 Concatenated	0.76	0.75	0.72	0.75
ViT Concatenated	0.76	0.76	0.73	0.76

Table 3.5: The results for accuracy metrics of Inception V3 trained on concatenated images vs. ViT trained on concatenated images.

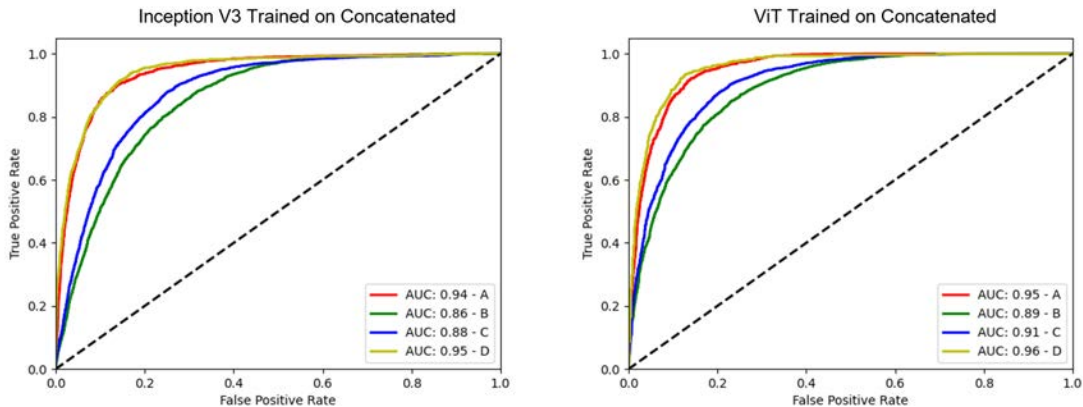


Figure 3.6: AUC/ROC curve for Inception V3 trained on concatenated CC images and the AUC/ROC curve for ViT trained on concatenated CC images

The results comparing Inception-V3 to ViT when using concatenated CC images shown in table 3.5 and figure 3.6. Using ViT with concatenated images provides on average a 1 percent increase to the test accuracy, macro F1, and weighted F1. The AUC/ROCs show that using ViT provided on average a 1 percent increase to the AUC of classes A and D, a 3 percent increase to class B AUC, and a 2 percent increase to class C AUC. Of all the changes to the performance metrics and AUCs, only the change to class C AUC was significant ($p = 0.0095$).

3.2.4 AdaBoosting Inception-V3 for Comparison to ViT

There was interest in the impact of the AdaBoost algorithm on the performance of a deep learning breast density classifier. Utilizing Skorch a PyTorch wrapper that allows for compatibility with sci-kit learn functions, a breast density classifier was developed using both an inception-based weak classifier (LIC) and AdaBoost. The LIC algorithm was modeled after Inception-V3, utilizing only 3 of the 9 inception modules that are present in Inception V3. This removal of inception modules was performed to "weaken" the classifier. The AdaBoost ensemble used with the LIC algorithm consisted of 15 classifiers. 5-fold cross-validation was performed using Inception V3 and the LIC algorithm. All tests used concatenated CC mammograms and training lasted for 100 epochs. The results from 5-fold cross-validation comparing the performance between Inception V3 and AdaBoosting of the LIC classifier are displayed below.

Inception-V3 vs LIC AdaBoost Algorithm			
Model	Test Accuracy	Macro F1	Weighted F1
Inception V3	0.75	0.72	0.75
ViT	0.76	0.73	0.76
AdaBoosting LIC	0.75	0.72	0.75

Table 3.6: The results for Inception V3 vs. AdaBoosting of the LIC classifier

Table 3.6 and figure 3.7 outline the 5-fold cross-validation results for AdaBoosting the LIC classifier compared to Inception-V3 without AdaBoost. Using AdaBoost with the inception-based weak classifier (LIC) did not provide any change to the test accuracy, macro F1, or weighted F1. The figure above shows the AUC/ROC for AdaBoosting the LIC classifier compared to the AUC/ROC for Inception V3. AdaBoosting the LIC classifier on average provided a decrease of 1 percent in the AUC for class A but a 3 percent increase in the AUC for class B and

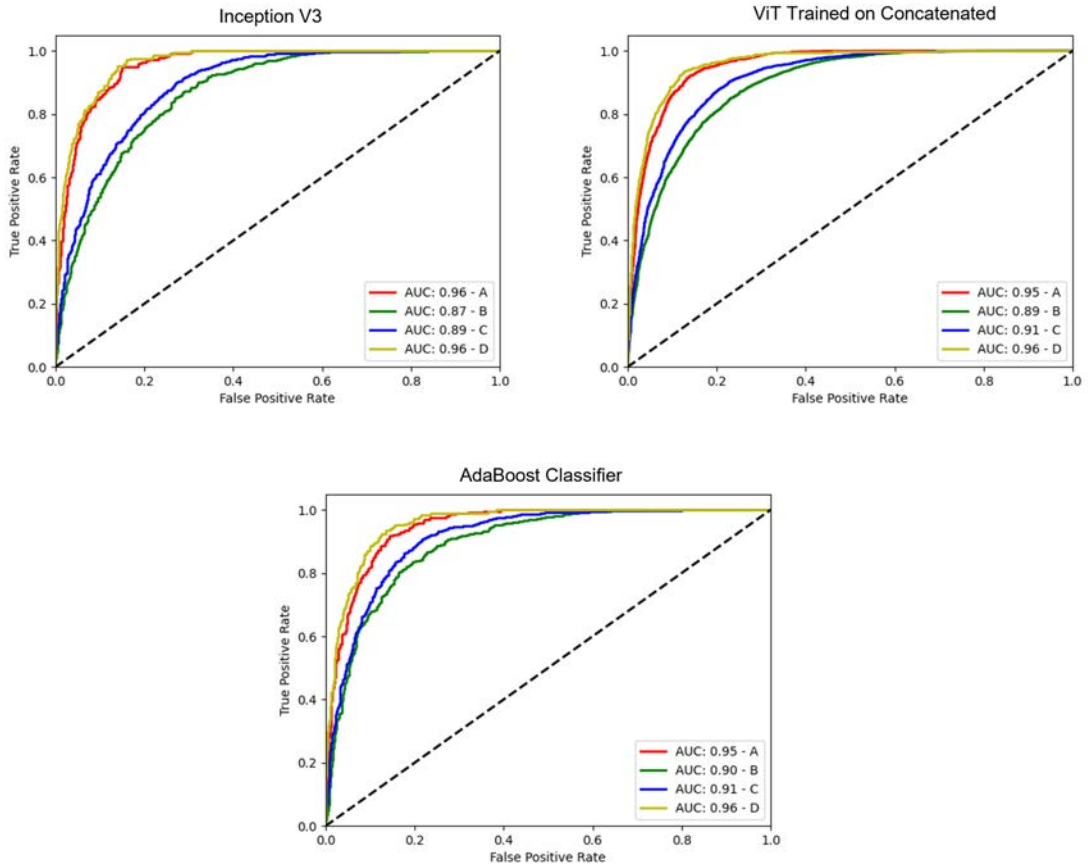


Figure 3.7: AUC/ROC curves for Inception V3, ViT, and the AdaBoosted LIC Classifier

a 2 percent increase in the AUC for class C. Of these increases, only the increase to class C AUC was considered significant ($p = 0.0095$). Comparing the results between LIC and ViT showed no statistically significant differences between the performance metrics evaluated.

3.2.5 Use of AdaBoost to Identify Noisy Labels

While AdaBoosting the LIC algorithm did not provide any statistically significant benefit to the performance of the classifier, it was hypothesized that the use of the AdaBoost algorithm could be used to identify and clean the noisy labels in the data set. Due to the subjective nature of breast density classification by a radiolo-

gist, there can be a degree of noise introduced into the data set. AdaBoosting has the tendency to overfit the noisy labels in a data set due to misclassified images possessing more weight than correctly classified images. If these images are those that possess noisy ground truth labels, one may be able to use AdaBoost to identify these images. During the training of the LIC algorithm using AdaBoost, the weights of the each the samples in the training set were recorded. The expectation is that samples with the highest weights at the end of training all of the weak classifiers in the ensemble will be images that are misclassified the most over the entire training procedure. After training, the final sample weights over 5 folds of the concatenated mammogram data set were sorted and list of the 100 patients with the highest sample weights at the end of the training procedure over these 5 folds was compiled. An expert radiologist (RWM) then reviewed the images from these 100 patients to determine the degree of noise of noise contained within these 100 patients. If the AdaBoost algorithm is increasing the weights of the noisy labels there should be a significant amount of relabeled images. The class transition counts and percentages are displayed in table 3.7.

Class Transition Count				
	Revised			
Initial	A	B	C	D
A	13	11	0	0
B	2	4	18	2
C	9	15	4	5
D	2	2	6	7

Class Transition Percentage				
	Revised			
Initial	A	B	C	D
A	54.17	45.83	0.0	0.0
B	7.69	15.38	69.23	7.69
C	27.27	45.45	12.12	15.15
D	11.76	11.76	35.29	41.18

Table 3.7: Class transition count and class transition percentage after radiologist re-evaluation of 100 "hard" images

Of the 100 patients, 72 (0.73% of the total 9899 patients) were assigned a new ground truth label after revision by the expert radiologist. After adjusting the data sets to account for the change in labels, another model of Inception V3 was trained using 5-fold cross-validation to determine if revision of these labels would provide any impact. The expectation was that the impact of this revision would not be enough to significantly impact the performance of the classifier due to the

small number of patients used in this experiment. The results of this test with the data set with relabeled patients against the initial patient labels is outlined in table 3.8.

Inception V3 Initial Labels vs Inception V3 Revised Labels				
Model	Validation Accuracy	Test Accuracy	Macro F1	Weighted F1
Inception V3 Initial	0.76	0.75	0.72	0.76
Inception V3 Revised	0.77	0.76	0.73	0.76

Table 3.8: The results for accuracy metrics of Inception V3 trained on concatenated images vs. ViT trained on concatenated images.

Adjusting the labels resulted in a 1 percent increase to the validation accuracy, test accuracy, and macro F1 scores. There was no change to the weighted F1 score. While there was no change, the fact that 72% of the 100 patients selected using AdaBoost shows that tracking the weights of samples while using the AdaBoost algorithm could allow for more efficient artificial intelligence assisted data cleaning. While this data cleaning could serve as an efficient means to identify samples in the dataset that may be noisy, too much of this cleaning could align the thresholds to the radiologist revising the labels.

CHAPTER 4

BREAST DENSITY SEGMENTATION USING DEEPLABV3

While the use of neural networks for classification of mammogram images has been shown to be a viable solution for automated approaches seeking to reduce inter-reviewer variability between radiologists, one could argue that breast density classification performed by a radiologist is not solely a classification task but also a segmentation task. Radiologist’s will either intentionally or subconsciously segment the dense tissue from the fatty tissue of the breast while performing their breast density assessment. It is for this reason that there was interest in the development of a deep-learning-based breast density segmentation algorithm that would not only provide visual dense tissue segmentation but also provide radiologists with a scaling a class probability system. In this chapter of my thesis, I explore the use of DeepLabV3 for semantic segmentation of dense breast tissue to assist in the standardization of radiologist assessment of breast density.

4.1 Imaging Data

Similar to previous experiments performed, the segmentation experiments in this thesis utilized mammogram images from only the CC image orientation. The use of exclusively the CC mammogram images from the C-View imaging format was also done, as per recommendation by an expert radiologist (RWM), to reduce the amount of error in the development of the segmentation algorithm. These two characteristics resulted in the identification of 37,284 CC/C-View images belonging to 17,625 patients from the larger data set of Mayo Clinic mammogram images. This subset of 37,284 images was then used to develop the data used in the creation

of the segmentation algorithm.

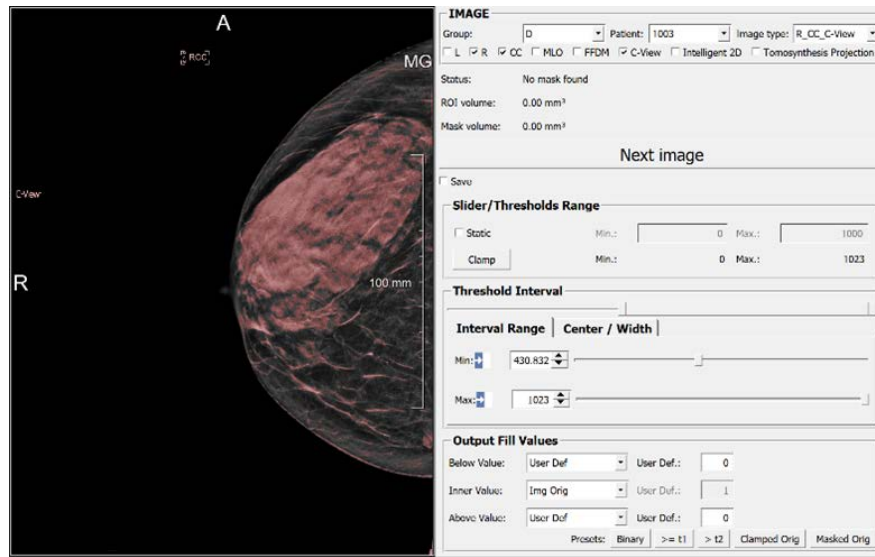


Figure 4.1: MeVisLab GUI as it was presented to the radiologist for segmentation mask development. The GUI utilized a threshold based slider to highlight the pixels to be included in the mask. These pixels were highlighted in red.

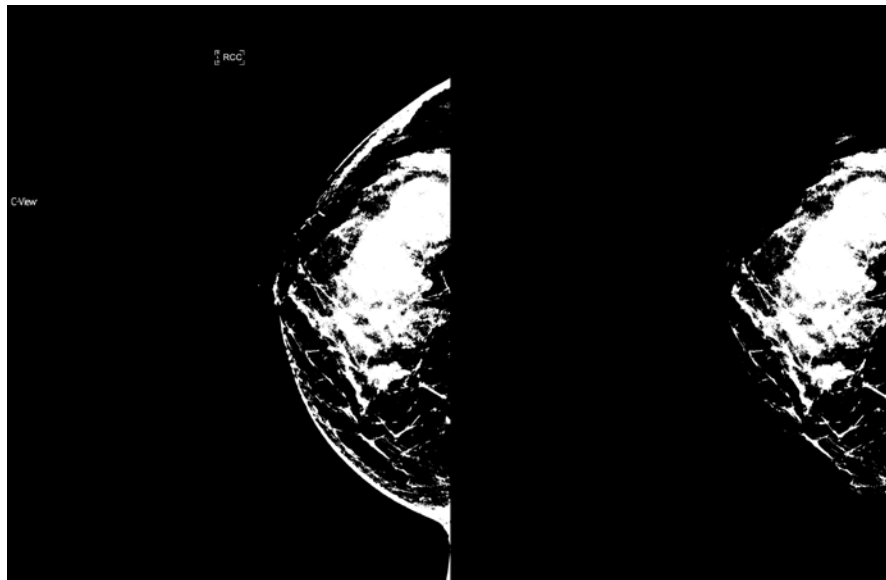


Figure 4.2: The use of a threshold-based slider has the tendency to include undesired regions of the mammogram in the mask. Regions of the mammogram that were undesired such as the radiographic labels and regions of high exposure from the curvature of the breast were manually removed.

4.2 Segmentation Development Set

Using MeVisLab (Fraunhofer MeVisLab medical Solutions AG, Bremer, Germany <https://www.mevislab.de/>), a Graphical User Interface (GUI) was used to present a patient’s mammogram images to an expert radiologist (RWM) (Figure 4.1). This radiologist then screened the mammogram images for proper breast positioning, proper exposure, and a lack of any unwanted artifacts. On images that met the radiologist’s screening criteria, a slider was used to adjust the pixel intensity threshold used to segment the dense pixels from the non-dense pixels. Once the radiologist was satisfied with the visual coverage of the segmented pixels, they then saved the segmented pixel information as a segmentation mask. While this method allowed for an acceptable segmentation of the dense tissue of the breast, pixels that are of similar intensity values to the dense tissue pixels are commonly included in the mask. Due to this coarse segmentation process, the masks were further refined using a script written using Python and OpenCV[20], a library consisting of hundreds of computer vision algorithms. This script was used to manually remove any included radiographic labels as well as high intensity pixels along the curvature of the breast. This process resulted in a data set of 688 expert verified images from 329 patients. Each patient possessed at least one left CC image (LCC) and one right CC image (RCC). To ensure that there is proper coverage of the breast tissue, some patients did possess more than one image for either one or both breasts. 5 folds of this set of 329 patients were created using and each fold was split in a 3-1-1 ratio to ensure that any variations in the performance was noted. Using dice scores to compare the algorithm output the radiologist’s (RWM) ground truth segmentations I can ascertain the performance of the algorithm and quality of the trained algorithm’s segmentations [13].

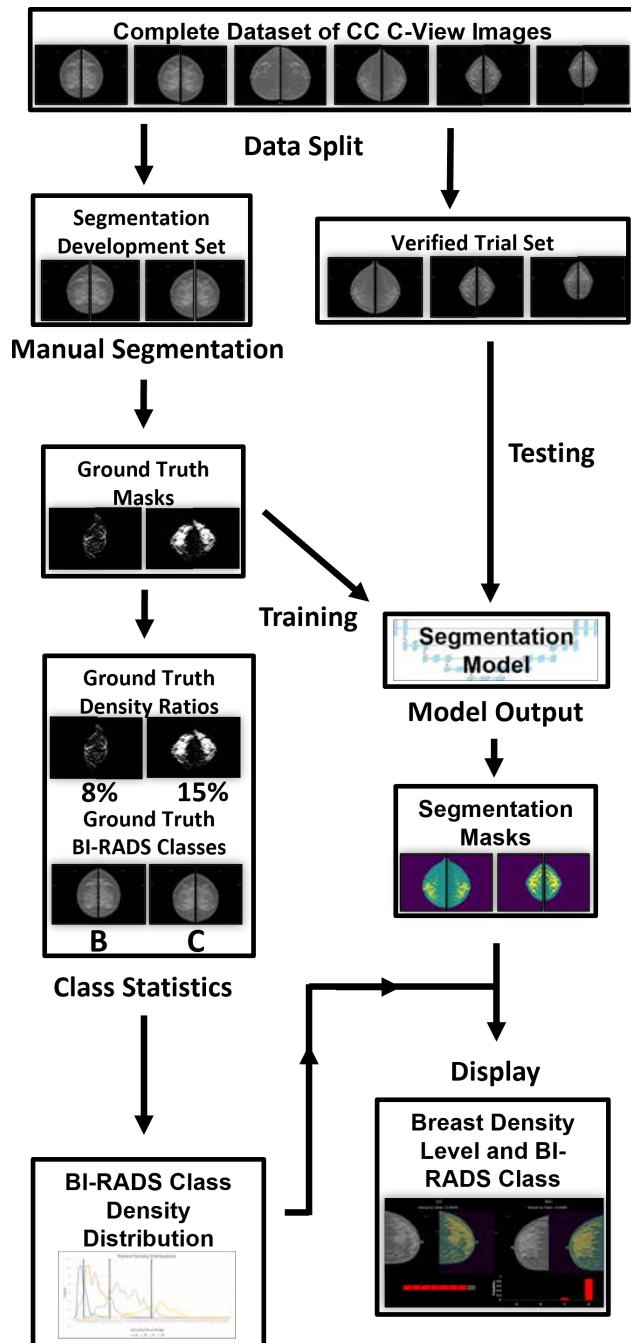


Figure 4.3: Pipeline for the development and evaluation of the breast-density segmentation algorithm. Segmentation training, validation, and test sets were created using 688 images from 329 patients from a total data set of 37,284 images from 17,625 patients. The density metric thresholds were extracted from the Segmentation Development Set. After training a segmentation model, the algorithm was applied to a verified set of 3,205 images from 1,522 patients. Applying the extracted thresholds to the density metric output by the model on the verified trial set, I determined the accuracy, a probability distribution, and a population distribution. The resulting data can then be used to display linear and probability scales to the radiologist using the automated tool.

4.3 Verified Trial Set

The creation of a “clean” Verified Trial Set (hold-out set) of 3,205 images from 1,522 patients was carried out. This was accomplished by examining both the mammogram images and the overlaid corresponding algorithm-produced segmentation mask for proper breast positioning, absence of artifacts, and adequate radiographic exposure. The goal of using this set was to assess segmentation-algorithm performance on a set of data resembling the training data. The resulting accuracy of the algorithm’s performance would represent a “best case” deployment scenario.

4.4 Calculation of Density Ratio from Segmentation Masks

The DeepLabV3 segmentation algorithm used in this study outputs a segmentation mask of an input mammogram image. This segmentation mask consists of pixels classified as one of three classes (background, fatty breast tissue, or dense breast tissue). From each of these output segmentation masks, the number of pixels in the fatty breast tissue and dense breast tissue classes was calculated. Adding these two values together results in the total breast tissue. With the three-pixel counts, calculation of the dense tissue to total tissue metric can be carried out. Figure 4.4 shows a sample CC mammogram image and its corresponding segmentation map.

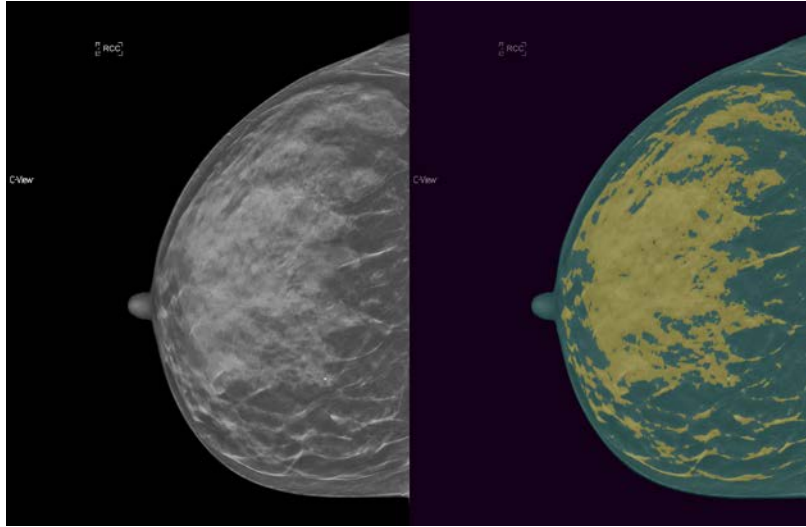


Figure 4.4: Original patient RCC mammogram image (left) with its corresponding segmentation map overlaid (right). Segmented dense tissue (yellow) and the segmented breast tissue (blue green) can be used to calculate the patients breast density for each of their breasts. Using this value from both breasts, the patient breast density can be determined. Segmentation maps can be displayed to the radiologist along with their placement on the linear scale and probability charts.

4.5 Threshold Determination

The proposed approach seeks to provide a linear scale of the aforementioned breast density metric. In order provide a scale with clear thresholds between each BI-RADS class to the radiologist, the Segmentation Development Set was used to determine the placement of these thresholds. Using a Python script, the best placement of these thresholds optimizing for BI-RADS classification accuracy with the radiologist created segmentation masks was performed. These thresholds can then be applied to the verified trial set to visualize the population distributions with thresholds along the linear scale of densities as well as determine a resulting accuracy.

4.6 Probability Distribution Methodology

The development of a distribution of BI-RADS class probabilities along the linear density scale of the verified trail set has the potential to provide radiologists with additional information during a breast density assessment. To develop this distribution, the probability of each class at a given density value was calculated. Using Sci-Kit Learn we can then use polynomial regression to develop a distribution of the class probabilities for each density metric. This probability distribution is visualized in figure 4.6. This methodology can also be used to determine the percent probability for each class and presented as an additional quantitative metric to the radiologist.

4.7 Results

Results outlining the dice coefficients after training the DeepLabV3 breast density segmentation algorithm on 5 folds of the segmentation data set are displayed in Table 4.1. The average dice coefficient for breast segmentation across the 5 folds was determined to 0.996 while the average dense segmentation dice coefficient was determined to be 0.726. The breast dice refers to the dice coefficient for segmenting the total breast tissue from the black background and radiographic labels. The almost perfect dice scores for breast dice show that this is a simpler task for the model to accomplish. The dense dice, which refers to the dice coefficient for the segmentation of the dense tissue from the fatty tissue within the breast, shows that the task of dense tissue segmentation is a more challenging task for the model. This is due to the variation in the pixel intensities being the means in which the dense tissue is segmented. There could be a number of factors that influence this pixel intensity such as the breast position, tissue thickness, or the imaging device used.

Following the training and evaluation of the segmentation algorithm perfor-

Table 4.1: Dice Coefficients Across 5-Folds of Segmentation Set

Dice Scores		
Fold	Breast Dice	Dense Dice
CV1	0.994	0.732
CV2	0.996	0.730
CV3	0.997	0.712
CV4	0.996	0.713
CV5	0.997	0.744
Average	0.996	0.726

mance in regard to dice coefficients, the segmentation algorithm was applied to each of the patient images in the verified trial set. The model outputs a segmentation map of the input image. We calculated the density metric for each of the output segmentation maps and saved the values to a separate file. This file was then used to determine the average accuracy, linear kappa, and quadratic kappa for BI-RADS classification when using the thresholds determined from the ground truth segmentations.

Table 4.2: BI-RADS Classification Accuracy Using Thresholds

BI-RADS Classification Accuracy Using Thresholds	
Accuracy	0.731
Linear Kappa	0.619
Quadratic Kappa	0.829

Table 4.2 displays the results for classification using the calculated density metric and thresholds determined from the segmentation development set. The accuracy achieved on BI-RADS classification using thresholds was 73.1%. The linear kappa score for classification showed a substantial level of agreement at 0.619 between our algorithm and radiologists while the quadratic kappa score shows a much higher agreement of 0.829. Using the calculated density metrics for each patient, we can determine the distributions for each BI-RADS class. We also used Sci-kit Learn to implement Polynomial Regression as a means to smooth the curves for better visualization. Using these patient-based density distributions (Figure 4.5), we can then derive probability curves (Figure 4.6) for each BI-RADS

class corresponding to each density value. Both Figure 4.5 and Figure 4.6 show the thresholds calculated from the Segmentation Development set placed along the distributions as dashed black line. Both the distribution of patients and the probability curves, along with the calculated thresholds (Table 4.3), can be used to provide additional visual tools to the radiologist at the time of a breast density screening.

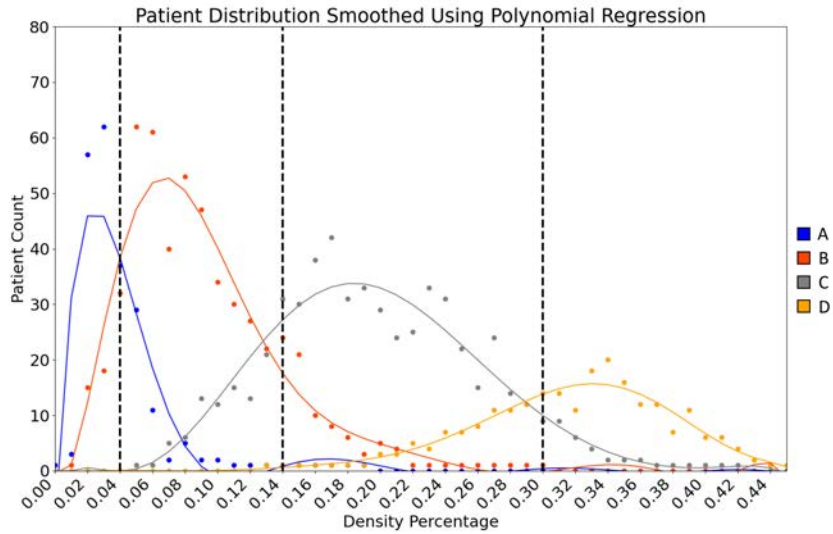


Figure 4.5: Patient based breast density distributions. Black dashed lines show the thresholds extracted from the segmentation development set. Clearer separation is exhibited between classes B and C. Class A and B exhibit more overlap in their distributions. This observation is maintained with classes C and D. Due to legislative guidelines, the distribution lends some insight into radiologist adjudication around BI-RADS class thresholds. Radiologists may be more confident or careful around the B/C thresholds than around the A/B and C/D thresholds.

Table 4.3: Density Thresholds

Density Thresholds	
Threshold	Density Metric
A/B	0.04
B/C	0.14
C/D	0.30

Placing the thresholds calculated from the Segmentation Development Set along the patient density distribution shows where the class thresholds lie. Us-

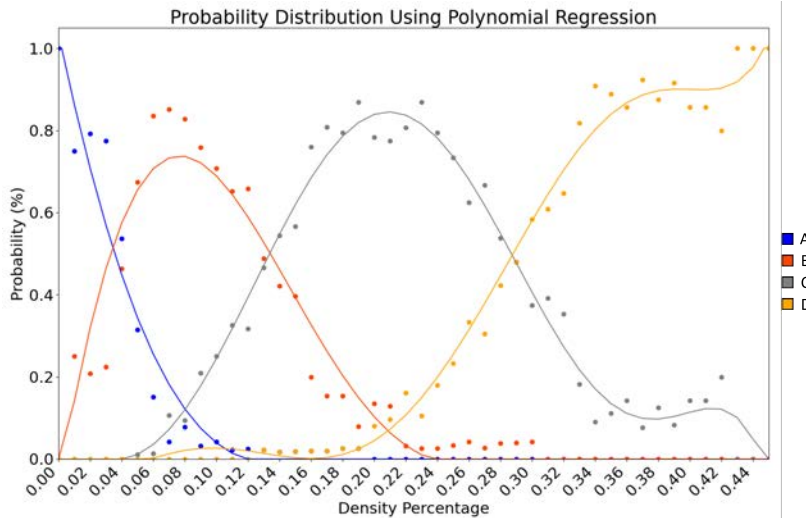


Figure 4.6: BI-RADS class probabilities based on the patient breast density percentages calculated from segmentation maps. The probability curves are determined using a polynomial regression model developed using the patient density distributions. Using this probability curve, we can output class probabilities for a given patient density to the radiologist.

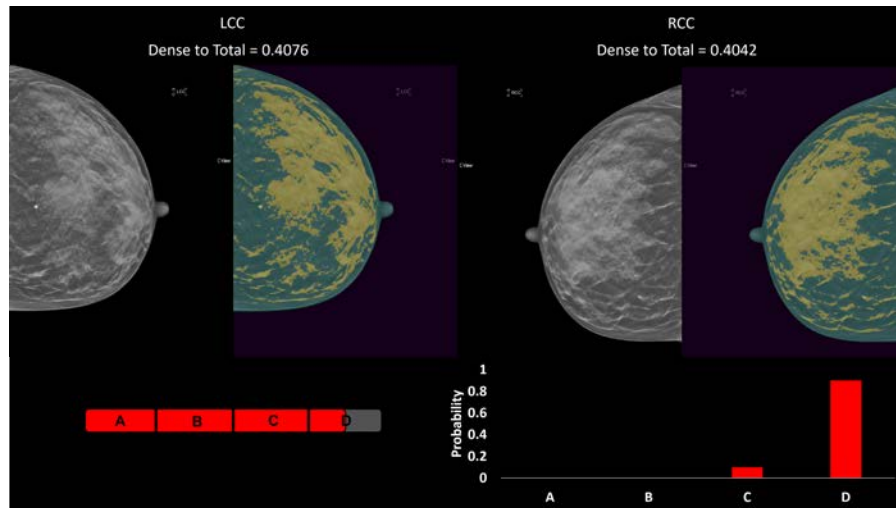


Figure 4.7: Sample display of the functionality of the semantic breast density segmentation algorithm. Segmentation maps can be overlaid over the original images and both quantitative metrics for each breast as well as a BI-RADS class assignment and class probabilities can be displayed to the user.

ing the segmentation output, thresholds, the distribution, and the probabilities, we can provide multiple visual tools to the radiologist at the time of the breast density screening. Figure 4.7 shows a sample of the visual tools that can be used

by the radiologist. Using the segmentation output we can overlay the segmentation mask over the original image to allow the radiologist to "approve" or "ignore" the segmentation algorithm's output. If the segmentation mask is satisfactory the radiologist can choose to utilize the linear density metric bar (shown on the lower left of Figure 4.7). The radiologist also has the option to visualize the class probability for that patient using the probability graph (shown on the lower right of Figure 4.7).

CHAPTER 5

CONCLUSIONS AND FUTURE WORK

For this thesis, I explored the use of deep learning algorithms to assist in the classification and segmentation of patient breast density. Work began with the development of mammogram image pre-processing technique that concatenated two mammogram images together. This concatenation resulted in better performance of the deep learning algorithms trained using these images. While exploring the impact of image concatenation, this thesis also investigated the implementation of a ViT as the model architecture used as the breast density classifier and compared the use of this architecture to the more mainstream CNN architecture. Following this work, the exploration of AdaBoosting a deep learning breast density classifier was explored. While this boosting technique did not result in any significant performance improvements. However, this boosting technique did appear to provide a means to identify and clean the noisy labels present in the large mammogram data set used in this thesis. While useful in identifying the noisy labels in the data set, revision of too many of the samples in the data set may just align the samples in the data set to the revising radiologist. These noisy samples present in radiological data set are a result of inter-rater variability due to the subjective assessment by individual radiologists. Rather than simply align the data and thresholds to an individual radiologist, there is significance in the reduction of the variability between radiologists to prevent noise in the data. To reduce this variability between radiologists, I propose that the assessment of breast density is inherently a segmentation task and that the development of a quantitative scoring system using semantic segmentation could aid in standardizing radiologist assess-

ments. The work performed using DeepLabV3 for semantic segmentation of breast density shows that this architecture can be used to develop a quantitative scoring system that also provides a linear and probability scale.

5.1 Future Work

It would be imprudent to believe that the work performed in this thesis is not in need of future work. The most immediate area of improvement lies in the inclusion of more image types and formats into the development of the breast density segmentation algorithm. The work performed in this thesis only considers the C-View images from the CC image orientation in the development of the segmentation algorithm. While this constraint was imposed in the essence of preserving the time spent by the expert radiologist creating ground truth segmentation masks, it can be viewed as a limitation of the work performed. The inclusion of more image types and formats would aid in the development of a more general model that is not restricted by the format of the input images. Another area of future work could investigate post deployment of the algorithm discussed in this thesis. The statistical analysis of inter-rate variability before and after the deployment of this algorithm in a hospital setting has the potential to expose the benefits of using such a system for assessment standardization. Future work could also focus on comparing the performance of the weakened Inception model used in the AdaBoost experiments against an AdaBoosted full Inception-V3 model trained for less epochs. The cleaning of more samples could also be performed to determine if the trend in class transitions is observed when more samples are present.

REFERENCES

- [1] ACR statement on reporting breast density in mammography reports and patient summaries. <https://www.acr.org/Advocacy-and-Economics/ACR-Position-Statements/Reporting-Breast-Density>. Accessed: 2022-11-10.
- [2] Densebreast-info, inc. legislative information. Available online at: <https://densebreast-info.org/legislative-information>, last accessed on Accessed: 2023-1-2.
- [3] Densitas. Available online at: <https://densitashealth.ca/solutions/density/>, last accessed on Accessed: 2023-1-5.
- [4] Powerlook density assessment. Available online at: <https://www.icadmed.com/powerlook-density-assessment.html>, last accessed on Accessed: 2023-1-5.
- [5] Synthesized 2d mammographic imaging theory and clinical performance. Available online at: <https://www.hologic.com/sites/default/files/C-View%20White%20Paper%2C%20Dr.%20Andrew%20Smith.pdf>, last accessed on Accessed: 2023-1-5.
- [6] AREFAN, D., TALEBPOUR, A., AHMADINEJAD, N., AND KAMALI-ASL, A. Automatic breast density classification using neural network. Journal of Instrumentation 10 (12 2015), T12002–T12002.

- [7] BADRINARAYANAN, V., KENDALL, A., AND CIPOLLA, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence 39, 12 (2017), 2481–2495.
- [8] BASSETT, L., HIRBAWI, I., DEBRUHL, N., AND HAYES, M. Mammographic positioning: evaluation from the view box. Radiology 188, 3 (1993), 803–806.
- [9] BYNG, J. W., BOYD, N. F., FISHELL, E., JONG, R. A., AND YAFFE, M. J. The quantitative analysis of mammographic densities. Phys. Med. Biol. 39, 10 (Oct. 1994), 1629–1638.
- [10] CHEN, L.-C., PAPANDREOU, G., SCHROFF, F., AND ADAM, H. Rethinking atrous convolution for semantic image segmentation, 2017.
- [11] CIATTO, S., BERNARDI, D., CALABRESE, M., DURANDO, M., GENTILINI, M. A., MARISCOTTI, G., MONETTI, F., MORICONI, E., PESCE, B., ROSELLI, A., STEVANIN, C., TAPPARELLI, M., AND HOUSSAMI, N. A first evaluation of breast radiological density assessment by QUANTRA software as compared to visual classification. Breast 21, 4 (Aug. 2012), 503–506.
- [12] DESTOUNIS, S., ARIENO, A., MORGAN, R., ROBERTS, C., AND CHAN, A. Qualitative versus quantitative mammographic breast density assessment: Applications for the US and abroad. Diagnostics (Basel) 7, 2 (May 2017), 30.
- [13] DICE, L. R. Measures of the amount of ecologic association between species. Ecology 26, 3 (July 1945), 297–302.
- [14] D’ORSI, C., SICKLES, E. A., MENDELSON, E. B., AND MORRIS, E. A. Breast Imaging Reporting and Data System: ACR BI-RADS breast imaging atlas. American College of Radiology, Reston, Va, 2013.

- [15] DOSOVITSKIY, A., BEYER, L., KOLESNIKOV, A., WEISSENBORN, D., ZHAI, X., UNTERTHINER, T., DEGHANI, M., MINDERER, M., HEIGOLD, G., GELLY, S., USZKOREIT, J., AND HOULSBY, N. An image is worth 16x16 words: Transformers for image recognition at scale, 2020.
- [16] D’ORSI, C., BASSETT, L., FEIG, S., ET AL. Breast imaging reporting and data system (bi-rads). Breast imaging atlas, 4th edn. American College of Radiology, Reston (2018).
- [17] ERIKSSON, M., CZENE, K., PAWITAN, Y., LEIFLAND, K., DARABI, H., AND HALL, P. A clinical model for identifying the short-term risk of breast cancer. Breast Cancer Res. 19, 1 (Dec. 2017).
- [18] HASTIE, T. J., ROSSET, S., ZHU, J., AND ZOU, H. Multi-class adaboost. Statistics and Its Interface 2 (2009), 349–360.
- [19] HELD, JR, L. I. SYMMETRY AND ASYMMETRY. In Quirks of Human Anatomy. Cambridge University Press, Cambridge, May 2009, pp. 17–32.
- [20] ITSEEZ. Open source computer vision library. <https://github.com/itseez/opencv>, 2015.
- [21] KELLER, B. M., CHEN, J., DAYE, D., CONANT, E. F., AND KONTOS, D. Preliminary evaluation of the publicly available laboratory for breast radio-density assessment (LIBRA) software tool: comparison of fully automated area and volumetric density measures in a case–control study with digital mammography. Breast Cancer Res. 17, 1 (Dec. 2015).
- [22] KRIZHEVSKY, A., SUTSKEVER, I., AND HINTON, G. E. Imagenet classification with deep convolutional neural networks. In Advances in Neural Information Processing Systems (2012), F. Pereira, C. Burges, L. Bottou, and K. Weinberger, Eds., vol. 25, Curran Associates, Inc.

- [23] LECUN, Y., BENGIO, Y., AND HINTON, G. Deep learning. Nature 521, 7553 (May 2015), 436–444.
- [24] LI, J., SZEKELY, L., ERIKSSON, L., HEDDSON, B., SUNDBOM, A., CZENE, K., HALL, P., AND HUMPHREYS, K. High-throughput mammographic-density measurement: a tool for risk prediction of breast cancer. Breast Cancer Res. 14, 4 (Aug. 2012).
- [25] LIN, T.-Y., MAIRE, M., BELONGIE, S., HAYS, J., PERONA, P., RAMANAN, D., DOLLÁR, P., AND ZITNICK, C. L. Microsoft coco: Common objects in context. In Computer Vision – ECCV 2014 (Cham, 2014), D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds., Springer International Publishing, pp. 740–755.
- [26] LIU, C.-C., TSAI, C.-Y., LIU, J., YU, C.-Y., AND YU, S.-S. A pectoral muscle segmentation algorithm for digital mammograms using otsu thresholding and multiple regression analysis. Computers and Mathematics with Applications 64, 5 (2012), 1100–1107. *Advanced Technologies in Computer, Consumer and Control.*
- [27] NG, K.-H., YIP, C.-H., AND TAIB, N. A. M. Standardisation of clinical breast-density measurement. Lancet Oncol. 13, 4 (Apr. 2012), 334–336.
- [28] OLIVER, A., TORTAJADA, M., LLADÓ, X., FREIXENET, J., GANAU, S., TORTAJADA, L., VILAGRAN, M., SENTÍS, M., AND MARTÍ, R. Breast density analysis using an automatic density segmentation algorithm. J. Digit. Imaging 28, 5 (Oct. 2015), 604–612.
- [29] OOMS, E. A., ZONDERLAND, H. M., EIJKEMANS, M. J. C., KRIEGE, M., MAHDAVIAN DELAVARY, B., BURGER, C. W., AND ANSINK, A. C. Mam-

- mography: Interobserver variability in breast density assessment. Breast 16, 6 (Dec. 2007), 568–576.
- [30] PAWAR, S., SAPATE, S., AND SHARMA, K. Machine learning approach towards mammographic breast density measurement for breast cancer risk prediction: An overview. SSRN Electron. J. (2020).
- [31] RONNEBERGER, O., FISCHER, P., AND BROX, T. U-net: Convolutional networks for biomedical image segmentation, 2015.
- [32] RUSSAKOVSKY, O., DENG, J., SU, H., KRAUSE, J., SATHEESH, S., MA, S., HUANG, Z., KARPATY, A., KHOSLA, A., BERNSTEIN, M., BERG, A. C., AND FEI-FEI, L. ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision (IJCV) 115, 3 (2015), 211–252.
- [33] SHELHAMER, E., LONG, J., AND DARRELL, T. Fully convolutional networks for semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence 39, 4 (2017), 640–651.
- [34] SIEGEL, R. L., MILLER, K. D., FUCHS, H. E., AND JEMAL, A. Cancer statistics, 2022. CA: A Cancer Journal for Clinicians 72, 1 (2022), 7–33.
- [35] SZEGEDY, C., LIU, W., JIA, Y., SERMANET, P., REED, S., ANGUELOV, D., ERHAN, D., VANHOUCKE, V., AND RABINOVICH, A. Going deeper with convolutions, 2014.
- [36] SZEGEDY, C., VANHOUCKE, V., IOFFE, S., SHLENS, J., AND WOJNA, Z. Rethinking the inception architecture for computer vision. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016), pp. 2818–2826.

- [37] USMAN, M., ZIA, T., AND TARIQ, A. Analyzing transfer learning of vision transformers for interpreting chest radiography. J. Digit. Imaging (July 2022).
- [38] WILD, C. P., WEIDERPASS, E., AND STEWART, B. W. World Cancer Report: Cancer Research for Cancer Prevention. Lyon, France, 2020.
- [39] WU, N., GERAS, K. J., SHEN, Y., SU, J., KIM, S. G., KIM, E., WOLFSON, S., MOY, L., AND CHO, K. Breast density classification with deep convolutional neural networks. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2018), pp. 6682–6686.

APPENDIX A

RELATED PUBLICATIONS, ACHIEVEMENTS, AND DELIVERABLES

C. Testagrose et al., "Impact of Concatenation of Digital Craniocaudal Mammography Images on a Deep-Learning Breast-Density Classifier Using Inception-V3 and ViT," 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Las Vegas, NV, USA, 2022, pp. 3399-3406, doi: 10.1109/BIBM55620.2022.9995206.

Placed 4th in the MICCAI Breast Density Federated Learning Challenge - Sponsored by NVIDIA

Code available at: <https://github.com/ctestagrose/Breast-Density-Classification>

APPENDIX B
ADDITIONAL RESULTS