Deep Learning for Clinical Mammography Screening

by

Nicholas Locascio

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of
Master of Engineering in Computer Science and Engineering
at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2017

© Nicholas Locascio. All rights reserved.

The author hereby grants to MIT permission to reproduce and to
distribute publicly paper and electronic copies of this thesis document
in whole or in part in any medium now known or hereafter created.

Author

Department of Electrical Engineering and Computer Science

June 10, 2017

Certified by

Dr. Regina Barzilay
Delta Electronics Professor, EECS Microsoft Faculty Fellow
Thesis Supervisor

Accepted by

Christopher J. Terman
Chairman, Master of Engineering Thesis Committee
Deep Learning for Clinical Mammography Screening

by Nicholas Locascio

Submitted to the Department of Electrical Engineering and Computer Science
in Partial Fulfillment of the Requirements for the Degree of
Master of Engineering in Electrical Engineering and Computer Science

Abstract

Breast cancer is the most common cancer among women worldwide. Today, the vast majority of breast cancers are diagnosed from screening mammography. Multiple randomized clinical studies have demonstrated that screening mammography can help reduce the number of deaths from breast cancer among women ages 40 to 74, especially for those over age 50 [4], and can provide women diagnosed with breast cancer more options for less aggressive treatment [7]. Screening mammography is the first entry into the funnel of clinical mammography. A screening mammogram can result in a suspicious finding, leading the patient to receive additional imaging, and even surgical biopsy if the additional imaging. Screening mammography, as the first part of this funnel, is a place for machine learning to have impact on the largest amount of patients. In this work, we apply machine learning models to tasks in clinical mammography such as density estimation, and Bi-Rads prediction.
Acknowledgments

I would like to thank my parents for their love, support, and guidance. I would like to thank my advisor, Regina Barzilay, for without, none of this would have been possible. I would like to thank Kyle Swanson, for his incredible willpower, drive, and passion. I would like to thank Anne Hunter for all the times she helped and guided me these past 5 years, and especially for her help in making 6.S191 possible.
Contents

1 Introduction 9

2 Background 13
   2.1 Clinical Mammography ................................. 13
      2.1.1 Breast Density ................................. 13
   2.2 Bi-Rads Assessment .................................. 14

3 Prior Work 17
   3.1 Bi-Rads Prediction ................................. 17
      3.1.1 Multi-View Full Resolution ...................... 17
   3.2 Density Estimation ................................. 17
      3.2.1 SVM ROI Density .............................. 17
      3.2.2 Deep Learning Based Density ..................... 18

4 MIT-MGH Dataset 19
   4.1 Cohort Statistics ................................. 19
   4.2 Image Preprocessing ............................... 20
      4.2.1 DICOM Medical File Format .................... 20
      4.2.2 Normalization and Squaring of the Images ...... 20

5 Deep Learning for Mammography Bi-Rads Assessment 21
   5.1 Challenges of high-resolution imagery ............... 21
   5.2 Method ............................................... 23
      5.2.1 Patch and Aggregation Model .................... 23
5.2.2 Patch Model ............................................. 24
5.3 Experiments .................................................. 25
  5.3.1 Training ................................................... 25
5.4 Patch Results ................................................ 26
5.5 Discussion ..................................................... 27
  5.5.1 Patch Model Predictions ................................. 27
  5.5.2 Patch Model Filter Visualizations ....................... 27
  5.5.3 Importance of Density: Conditional Patch Model .......... 29
  5.5.4 PatchNet Feature Representation Evaluation ............... 29
5.6 Aggregation Net ............................................... 30
  5.6.1 Multi-View, Multi-Time Aggregation ....................... 30
  5.6.2 Triage Task ............................................... 32
  5.6.3 Results .................................................. 32

6 Conclusion .................................................... 35
List of Figures

2-1 Breasts of various densities.  **A:** Almost entirely fat  **B:** Scattered fibro glandular densities  **C:** Heterogeneously dense  **D:** Extremely dense

2-2 Breasts of various BI-RADS categories.  **BI-RADS 0:** Needs Additional Evaluation  **BI-RADS 1:** Negative  **BI-RADS 2:** Benign Finding We group 0 vs 1 and 2 to pose the problem as callback vs no-callback.

5-1 Patch-Aggregation Model Setup.
5-2 Final AUC curve for PatchNet. (AUC=0.88)
5-3 Learning Curve for Patch AUC metric after each epoch.
5-4 Patch-Net Model Predictions.
5-5 Patches that maximally activate the 3rd conv filter of the 6th layer
5-6 Patches that maximally activate the 7th conv filter of the 6th layer
5-7 PatchNet with varying density input conditions
5-8 t-SNE Plot of learned patch features
5-9 Our model’s scores and threshold for where it triages patients.
Chapter 1

Introduction

Breast cancer is the most common cancer among women worldwide. Today, the vast majority of breast cancers are diagnosed from screening mammography. Multiple randomized clinical studies have demonstrated that screening mammography can help reduce the number of deaths from breast cancer among women ages 40 to 74, especially for those over age 50 [4], and can provide women diagnosed with breast cancer more options for less aggressive treatment [7]. Screening mammography is the first entry into the funnel of clinical mammography. A screening mammogram can result in a suspicious finding, leading the patient to receive additional imaging, and even surgical biopsy if the additional imaging. Screening mammography, as the first part of this funnel, is a place for machine learning to have impact on the largest amount of patients.

Despite significant recent advancements in computer vision and deep learning, all the screening mammograms today are still read manually.

Deep Learning techniques have revolutionized various fields such as object recognition, speech recognition, and machine translation. In this work, we apply deep learning techniques to mammogram images to automatically learn relevant and useful image representations to perform various prediction tasks. We show that deep learning can be used to detect subtle features in the image that are predictive of lesion malignancy.

Mammography imaging presents a unique challenge for image processing systems
because cancerous masses are small, subtle, and local features. Furthermore, mam-
mography images are extremely high-resolution, whereas much work in image pro-
cessing has focused on low-resolution images. Furthermore, proper modeling of tumor
growth requires not only viewing the current mammogram, but also viewing the prior
mammogram. Furthermore, as asymmetries between the breasts are an indicative.
Furthermore, sometimes a tumor is only visible in a specific scanning view orientation,
or the visibility of the tumor in one view but not the other can indicate other factors
of risk. In this way, to truly model the problem, a model must not only deal with a
single high-resolution image, but multiple concurrently. The processing of multiple
images concurrently in a single model only compounds the problems already faced
with high-resolution imagery.

We propose a solution to the various challenges associated with high-resolution
mammographic imagery that is able to process mammograms at full resolution with-
out resorting to destructive downsampling.

We train this model on a large proprietary database of mammograms, collected
in partnership with MGH. We apply this model to a variety of prediction tasks, like
Bi-Rads (whether the breast should be called back for additional imaging), and breast
density (a measure of fatty composition of the breast).

Our contributions consist of the following:

1. Achieve 91.2% Accuracy on Density Prediction, better than the human agree-
   ment accuracy of 88.6%.

2. Achieve 0.882 AUC on Patch-level Bi-Rad classification.

3. Demonstrate a system that can triage 35% of all mammograms into a ’clearly
   benign’ category with equal-to-human false negative rate. If applied in clinic,
   this would save 56,000 mammograms from having to be read by radiologists,
   allowing them to focus their time and energy on the hard cases where it is not
   clear.

4. Highlight portions of mammograms that lead to the most and least confident
   predictions in our CNN.
5. Demonstrate that learned local representations of high-resolution images are useful in global prediction tasks. This demonstration is a key insight that can be applied to other high-resolution medical imagery, like pathology reports which exist in the Giga-Pixel space, around 100x times larger than mammogram resolution.
Chapter 2

Background

2.1 Clinical Mammography

2.1.1 Breast Density

Breast Density is a four-tiered breast composition rating used by clinical radiologists as part of mammography screening guidelines. Density represents the proportion of different fatty and non-fatty tissues in the breast. The four tiers of breast density are as follows:

1. Almost entirely fat (less than 25 percent glandular).
2. Scattered fibroglandular densities (approximately 25-50 percent glandular).
3. Heterogeneously dense (approximately 51-75 percent glandular).
4. Extremely dense (more than 75 percent glandular).

Mammograms corresponding to these tiers are shown in Figure 2-1.

Breast density is an important measure in clinical mammography because it has been found to correlate with an increased chance of breast cancer. Furthermore, dense breasts often obscure masses and tumors, making accurate cancer detection difficult for the radiologist. Despite its importance in the assessment of a
Figure 2-1: Breasts of various densities. **A:** Almost entirely fat **B:** Scattered fibroglandular densities **C:** Heterogeneously dense **D:** Extremely dense

patient's risk, it is not a consistently evaluated metric, with radiologists often disagreeing on the density of the same breasts. For these various and many reasons, automated breast density is an important and high-impact area for machine learning to bring consistency, accuracy, and fairness to the clinical evaluation of a patient’s mammogram.

### 2.2 Bi-Rads Assessment

In screening mammography, radiologist determines a Bi-Rads Assessment for the patient. Bi-Rads assessments determine if a patient will be called back for additional imaging, often returning for an MRI of the suspect breast. There are three possible designations of Bi-Rads, **Bi-Rads 0:** Needs Additional Evaluation, **Bi-Rads 1:** Negative, and **Bi-Rads 2:** Benign Finding. Bi-Rads 0 will lead to a patient callback, where 1 and 2 will lead to no callback. The difference between Bi-Rads 1 and 2 is subtle and minimal. While both indicate no callback, Bi-Rads 2 indicates that there was a suspicious finding, but that it was ruled to be benign and non-cancerous. This score is often given for a patient if there is a suspicious region that had been biopsied in a previous year, with the result being benign. The suspicious region will still exist in the current scan because it still exists in the breast, but the radiologist knows that this region had already been determined to be benign, so will indicate its presence, but not recommend a patient callback. Bi-Rads values with corresponding examples
of breasts given the respective assessments are shown in Figure 2-2.

Figure 2-2: Breasts of various BI-RADS categories. **BI-RADS 0:** Needs Additional Imaging Evaluation **BI-RADS 1:** Negative **BI-RADS 2:** Benign Finding We group 0 vs 1 and 2 to pose the problem as callback vs no-callback.
Chapter 3

Prior Work

3.1 Bi-Rads Prediction

3.1.1 Multi-View Full Resolution

Geras et al [2] proposed a model which uses a CNN architecture with aggressive down-sampling to process all four mammographic views to produce a Bi-Rads assessment for the patient. Their model operates on the full pixel space which they are able to fit into their Inception architecture by reducing the number of filters of each convolution. In reducing the number of filters, they limit the power of their model. Furthermore, by operating on full pixel space of the global image, their training scheme suffers from the poor signal-to-ratio problem discussed earlier.

3.2 Density Estimation

3.2.1 SVM ROI Density

Subashini et al [6] trained a standard SVM model to automatically assess breast density. Their approach is limited to assess density only once given an ROI that does not contain noise, artifacts, or the pectoral muscle.
3.2.2 Deep Learning Based Density

Chul et al [1] trained a CNN to assess density from an entire mammogram. Though because their dataset is limited to just 400 images, their model cannot reliably learn useful image representations so they augment the image representation with the eigen-image of the mammogram which provides additional hand-crafted feature extraction.
Chapter 4

MIT-MGH Dataset

This is a Health Insurance Portability and Accountability (HIPAA)-compliant, retrospective study conducted within Partners Health-Care approved guidelines. Screening mammograms 76,897 accessions for 53,092 patients aged between 28 and 99, collected between 2012-2016 at Massachusetts General Hospital were used in this study.

4.1 Cohort Statistics

Our Cohort consists of 53,092 patients, with two sets of mammograms for every patient: the most recent mammogram and the prior mammogram. Prior mammograms must have been taken at least 365 days prior to the recent mammograms. We exclude patients with breast implants for this study. Of the 53,092 patients, 9.26% have almost entirely fat breasts (type 1), 46.35% have type scattered fibroglandular tissues (type 2), 39.13% have heterogeneously dense breasts (type 3), and 4.93% have extremely dense breasts (type 4). Of the accessions in the recent set, 5.5% have Bi-Rads 0 (callback), 68.4% have Bi-Rads 1, and 28.8% have Bi-Rads 2 assessments.
4.2 Image Preprocessing

4.2.1 DICOM Medical File Format

Mammograms, like many medical scans, are stored in the DICOM [5] file standard. DICOM files contain both the pixel values of the scan performed and also metadata about the scan, patient, and machine used to perform the scan. In order to facilitate our work in image processing, we extract the pixel values from the DICOM file and save them to a png file using the tool GDCM toolkit [[3]].

DICOM images are not stored in the standard 0-255 range of typical image formats. Rather, each of the various vendors of mammography scanning machines produces values of different ranges, and to properly represent and preserve the information created by the scan, it is important to take these non-uniform ranges into account. A few standard approaches exist for handling this scaling properly. We adopt the approach known as WindowWidth which uses a table of vendor-specified ranges of values to scale from. This preserves the proper range, and ensures that our automated algorithms will view the image at the same intensity and scaling that a radiologist views during an exam.

4.2.2 Normalization and Squaring of the Images

Mammogram images created from different vendors have different sizes, shapes, and ratios. When normalizing these images, it is important to retain the original aspect ratio of the pixels in the mammogram to avoid any stretching or warping which could destroy information. To normalize and handle the varying vendor standards properly, we perform square cropping and resizing of the images. Mammograms are taller than they are wide, so we can perform square cropping by making a square with side length equal to the original width and centering it when performing the crop. We lose some pixels in this process, but the pixels are most often not of the breast. Our resulting image is a square image of dimension 3328x3328. We use these squared images as the base for all of our experiments.
Chapter 5

Deep Learning for Mammography

Bi-Rads Assessment

5.1 Challenges of high-resolution imagery

Much of deep learning for image classification has dealt with classifying natural images in the resolution range of approximately 300x300 pixels. Typical digital mammogram images are on the order of 3000x3000 pixels - an order of magnitude larger than typical image classification scenario. This high-resolution imagery presents many-fold problems.

Memory Constraints

The first problem is that in order to train our model via backpropogation, it is necessary to convolve over the entire image, store the activations for each layer, compute the gradient over these activations with respect to the loss, and finally update our network weights by applying the gradient. This process is a atomic operation that requires all activations be stored in memory at once. Popular models like Inception, using a batch size of 128, require approximately 6 GB of GPU memory when run at 224x224 image resolution. However, when run on high-resolution, this same model requires approximately 50 GB of GPU memory. This presents an issue because the
largest commercially-available GPU only has 12 GB of memory. It is possible to partially alleviate this problem by lowering the batch size of SGD. With a batch-size of 1, we can fit our high-resolution images into GPU memory. However using such a small batch size introduces a large amount of noise into the update gradients, which prove problematic in achieving any learning.

**SGD and Signal to Noise Ratios**

Typical training of neural networks involves the use of stochastic gradient descent. SGD equations

\[
Q(w) = \frac{1}{n} \sum_{i=1}^{n} Q_i(w) \tag{5.1}
\]

\[
w := w - \eta \nabla Q(w) = w - \eta \sum_{i=1}^{n} \nabla Q_i(w) \tag{5.2}
\]

The important thing to note is that we are estimating the gradient sampling a limited sample from it, which means a noisy gradient. This usually does not present an issue, because the learning signal is strong enough such that it prevails over the noise introduced by SGD. However, in scenarios of high-resolution imagery, the signal to noise ratio of each individual image can extremely small. Some tasks that contain global image features have a strong enough signal because most pixels contribute to the classification. However, some tasks that are heavily reliant on extremely local image features contain signal small enough to be obscured by the noise introduced in SGD.

**Speed of Batch Iteration**

Performing convolutions across images that have 10x larger significantly increases the iteration time for training our model.
5.2 Method

In this section, we describe various approaches to estimating breast density via machine learning models. We treat the problem as a classic image classification problem.

5.2.1 Patch and Aggregation Model

We use a two-stage patch-extract-and-aggregate framework for extracting relevant visual features, and then aggregate those extracted features to build a predictive BI-RADS model. As previous work shows, building powerful BI-RADS models in the high resolution space of mammograms presents many difficulties: namely in (1) low signal to noise ratio, (2) slow training time, (3) limited capacity of model by memory constraints. We are able to side-step many of these issues with our patch-extract-and-aggregate framework.

The framework is as follows: our patch feature detector (PatchNet) is trained to discriminate small crops of the mammogram as either cancerous or benign. We then use this trained patch feature detector to produce feature maps for entire mammograms by running the patch model over the full image in a sliding-window fashion. This produces a feature map for the entire image that we then pass to our aggregation model (AggregationNet). This aggregation model operates over the extracted features from our PatchModel to produce a global breast-level prediction. A visual representation of our setup is shown in Figure 5-1.

Our approach solves many of the issues faced with the naive full-resolution global BI-RADS prediction approach which enables our model to learn powerful feature detectors relevant for cancer detection.

Sampling from these positive examples, and training on images where the majority of pixels contribute to the classification vastly improves our signal-to-noise ratio. Furthermore, these small images enable powerful full-fledged vision models like Inception and Resnet-50 without any modifications. We can train these architectures with large batch sizes and fast batch time due to the smaller image size.
5.2.2 Patch Model

We rely on a small subset of images (around 1,000 positive cases) where the tumor is explicitly annotated. The model we developed has two stages in the processing pipeline. First, we divide an image into small patches (corresponding to a grid division). All the patches that intersect with the annotated tumor are considered positive, the rest are considered negative. This data is used for training PatchNet, a Resnet-18 model that distinguishes between suspicious and normal patches. By training our
PatchNet to detect tumors, we learn a predictive feature representation for every patch. During the second stage, we aggregate these representations from which the model learns the BI-RAD.

**Patch Dataset Creation**

We asked a team of MGH radiologists to annotate regions of interest (ROI’s) for Bi-Rads 0 cases. These mammograms are annotated with the location rationale of the patient callback. These rationale are circles of the suspicious region of the mammogram. They annotated 1,241 ROI’s, with average radius of 212 pixels for each ROI.

Using these ROI’s we split each of the original full 1,241 images into a grid of 4,356 patches. This produces a very large dataset of 5,288,184 patches. We exclude 99% of patches that are pure black to reduce redundancy of our dataset and are left with 2,456,133 patches. We consider patches with 1/4 of overlap or more with the original ROI’s to be positives. All other patches are considered negative. This produces 19,856 positive patches and 2,436,268 negative patches.

**5.3 Experiments**

We split our patch dataset into train/dev/test (60/20/20). We utilize early stopping to achieve our best dev AUC, and we report the corresponding test AUC. We train our patch model for 130 epochs, using ADAM Optimizer with learning rate=0.001, and beta=0.99. We use BatchNormalization with momentum = 0.99.

**5.3.1 Training**

**Balanced Sampling**

Because the majority of the breast exists outside of the small region annotation, our patch dataset is heavily skewed toward negative cases (99.918%). Training a model using Stochastic Gradient Descent with such a large imbalance will result in a model
that never learns to discriminate positive from negative because positives are so rare in each batch. Instead of sampling randomly from the distribution as classic SGD requires, we sample from the positive and negative cases with equal representation, ensuring each batch consists of 50% of each class. This resolves the data skew issue and allows our model to learn feature detectors and an ultimate decision boundary.

5.4 Patch Results

We train our model from 1200 epochs and achieve AUC=0.882. Figure shows the corresponding ROC curve, and Figure 5-3 shows our learning curve.

![Figure 5-2: Final AUC curve for PatchNet. (AUC=0.88)](image)

Figure 5-2: Final AUC curve for PatchNet. (AUC=0.88)
5.5 Discussion

5.5.1 Patch Model Predictions

Because our patch model produces predictions at the patch level, we can run a sliding window of our classifier over an entire image to produce a heatmap of predictions. We visualize these predictions overlayed atop the original image in Figure 5-4.

5.5.2 Patch Model Filter Visualizations

We can also evaluate the quality of our patch predictor by analyzing the representations that it learns. We can visualize our model’s learned patch feature representations by analyzing the model’s filters. Figure 5-5 shows the patches that maximally activate the 7th filter in the 6th conv layer. This filter is detecting tissue-distortion abnormalities, often associated with breast cancer. We can see that different filters learn to detect different types of features. Figure 5-6 is a filter that detects benign
densities in the breast.

Figure 5-5: Patches that maximally activate the 3rd conv filter of the 6th layer.
5.5.3 Importance of Density: Conditional Patch Model

To evaluate the importance of density in predicting a suspicious region, we train a PatchNet to take additional feature of the patient’s prior density. With a PatchNet that produces predictions conditional on a density input, we can sample different densities to see how the model’s predictions change.

Notice that in Figure 5-7, across the board, high density inputs lead to more suspicious regions. This is due to the general correlation between high density and callback rates. As high density often obscures masses, radiologists have high recall rates for high dense breasts.

5.5.4 PatchNet Feature Representation Evaluation

PatchNet trained to detect local suspicious regions, however in learning to detect suspicious regions it should also learn disentangled feature representations for each patch. These features are length 512 vectors, so we dimensionality reduce them via T-sne visualize the reduced features in Figure 5-8.
5.6 Aggregation Net

Using our trained PatchNet which can discriminate local suspicious regions from benign regions, we move to the task of full-image classification. We slide our PatchNet over the full image, producing a feature map of 66x66x512. PatchNet produces a feature 512-length feature representation, which we separate into separate channels and arrange in a grid to preserve the spatial arrangement of features extracted from neighboring patches.

5.6.1 Multi-View, Multi-Time Aggregation

The feature representation provided by our PatchNet proves to be remarkably compact, which enables us to explore aggregation not only over a single high-resolution
image, but 8 high-resolution images. Radiologists often use 8 images in their assessment of a patient. There are four views per visit: R-MLO, L-MLO, R-CC, L-CC. Furthermore, radiologists additionally consider these four views from the patient’s last mammogram. Cancer growths manifest themselves as changes in the breast, and this temporal information is highly informative for a correct assessment.

We extract Patch Features for these eight images, to create 8 66x66x512 feature maps. Each feature map gets processed by a 18-layer Resnet-like architecture that produces a 1x128 feature map for that entire image. We then concatenate each of
these 8 1x128 feature maps into a single vector, and classify from this concatenated vector.

This 8-image setup allows us to incorporate the full context of the patient’s history and exams to provide the most accurate assessment.

5.6.2 Triage Task

Automated Patient Triage is a task we propose in this work that aims to improve clinical mammography care by improving the specificity of the human radiologists. Automated Patient Triage involves utilizing automated detection algorithms to prune away easy benign cases from hard, suspicious cases. We define easy cases by the following: if a breast is low-density and benign, it is easy. We define hard cases to be all other cases, in which either density is high, or the breast looks suspicious.

The purpose of the triage task is to allow radiologists to focus their time and energy on the difficult cases. At MGH, 10 radiologists perform 100,000 mammograms per year. The vast majority of these cases fall into the easy case, which creates unnecessary overhead and work for the radiologists. Furthermore, radiologists at an institution will vary in experience and skill. With an automated patient triage in place, the senior radiologists can perform the screening for the hard cases, and the junior radiologists can perform the screening for the easy cases.

5.6.3 Results

Our approach is able to triage 35% of patients into the 'safe' triaging category while retaining a false negative rate equal to human performance. If applied in clinic, this would save 56,000 mammograms from having to be read by radiologists, allowing them to focus their time and energy on the hard cases where it is not clear. Figure 5-9 shows the model’s scores and threshold for where it triages patients.
Figure 5-9: Our model’s scores and threshold for where it triages patients.
Chapter 6

Conclusion

High-Resolution medical imagery presents many difficult challenges for standard deep learning approaches. We present a model capable of overcoming many of these challenges to perform automated mammography triaging of 35% of patients. The ability to triage a large portion screening mammogram patients represents a significant potential to improve care and reduce costs in clinical mammography. Furthermore, our approach can be applied to other high-resolution medical images such as tomosynthesis scans.


