Mass classification of digital mammograms

using convolutional

neural networks

By

Elijah Franklin

A Thesis
Submitted to the Faculty of
Mississippi State University
in Partial Fulfillment of the Requirements
for the Degree of Master of Science
in Electrical Engineering
in the Department of Electrical and Computer Engineering

Mississippi State, Mississippi

May 2018
Mass classification of digital mammograms

using convolutional

neural networks

By

Elijah Franklin

Approved:

__________________________
John E. Ball
(Major Professor)

__________________________
Lori Bruce
(Committee Member)

__________________________
Bo Tang
(Committee Member)

__________________________
Jim Fowler
(Graduate Coordinator)

__________________________
Jason M. Keith
Dean
Bagley College of Engineering
This thesis explores the current deep learning (DL) approaches to computer aided diagnosis (CAD) of digital mammographic images and presents two novel designs for overcoming current obstacles endemic to the field, using convolutional neural networks (CNNs). The first method employed utilizes Bayesian statistics to perform decision level fusion from multiple images of an individual. The second method utilizes a new data pre-processing scheme to artificially expand the limited available training data and reduce model over-fitting.

Key words: Deep Learning, Machine Learning, Convolutional Neural Networks, Mammography
DEDICATION

To Allison and Teddy.
ACKNOWLEDGEMENTS

This work was supported by the Digital Database for Screening Mammography for the compilation of case files.

The findings and opinions in this thesis belong solely to the author, and are not necessarily those of any contributor.

I thank my committee for their comments on this thesis, and I thank John E. Ball for directing this research as well as providing access to his personal deep learning facilities.
# TABLE OF CONTENTS

DEDICATION ................................................................. ii

ACKNOWLEDGEMENTS ....................................................... iii

LIST OF TABLES ............................................................. vii

LIST OF FIGURES .......................................................... viii

CHAPTER

1. INTRODUCTION ........................................................... 1

   1.1 Breast Cancer Impact .............................................. 1

   1.2 Current State of Breast Cancer Diagnosis and Prognosis ........ 2

      1.2.1 Local Treatments ............................................ 4

      1.2.2 Systemic Treatments ......................................... 5

   1.3 Mass Morphology .................................................... 5

   1.4 Machine Learning Overview ....................................... 7

   1.5 Applying Deep Learning to Medicine ............................ 9

   1.6 Contributions ..................................................... 11

2. BACKGROUND ............................................................ 12

   2.1 Convolution Neural Network Building Blocks .................. 12

      2.1.1 Convolution .................................................. 12

      2.1.2 Activation Techniques ...................................... 13

      2.1.3 Pooling ....................................................... 15

      2.1.4 Fully Connected and Softmax ............................... 15

      2.1.5 Optimization ................................................ 16

      2.1.6 Generalization Techniques ................................. 16

   2.2 Convolutional Neural Network Evolution ......................... 17

      2.2.1 AlexNet ....................................................... 17

      2.2.2 Inception ..................................................... 18

      2.2.3 VGG Net ...................................................... 18

      2.2.4 Residual Networks .......................................... 19
2.2.5 Network Architecture Influence ........................................... 20
2.3 Digital Database Screening Mammography ............................ 21
2.4 Data Augmentation ............................................................. 21
2.5 Data Fusion ................................................................. 23
2.6 Bayesian Statistics ........................................................... 24
2.7 Receiver Operating Characteristic ...................................... 24
2.8 CAD Mammography .......................................................... 25
2.8.1 Shallow ................................................................. 25
2.8.2 Deep Learning ............................................................ 27
2.8.2.1 Deep vs. Shallow ............................................... 29
2.8.2.2 Unsupervised Approaches ................................. 30
2.8.2.3 Supervised Approaches ....................................... 30

3. UNREGISTERED MULTI-VIEW MAMMOGRAPHIC IMAGE CLASSIFICATION USING SELF-NORMALIZING CONVOLUTIONAL NEURAL NETWORKS ........................................... 32

3.1 Abstract ........................................................................... 32
3.2 Introduction ..................................................................... 33
3.2.1 Digital Mammography .............................................. 33
3.2.2 Deep Learning Mammography ................................... 34
3.2.3 Self Normalizing Neural Networks ............................... 36
3.3 Data ............................................................................. 37
3.3.1 Digital Database for Screening Mammography ............ 37
3.3.2 Seed Point Acquisition ............................................. 41
3.3.3 Pre-processing Steps .................................................. 42
3.3.4 Augmentation ............................................................. 43
3.3.5 Down-sampling .......................................................... 44
3.3.6 Gaussian Filtering ..................................................... 45
3.3.7 Non-linear Stretching ................................................ 47
3.3.8 Data Input Structure ................................................... 49
3.4 Methodology ................................................................... 50
3.4.1 Neural Network Configuration .................................. 52
3.4.2 Experiment Schedule ................................................ 53
3.5 Results .......................................................................... 54
3.5.1 Channel Manipulation ................................................ 54
3.5.2 Activation ................................................................. 55
3.6 Discussion ....................................................................... 56
3.6.1 Channel Manipulation ................................................ 57
3.6.2 Activation ................................................................. 58
3.7 Contributions and Future Work ....................................... 59
4. RAPID SCREENING MAMMOGRAPHIC ANALYSIS USING CONVOLUTIONAL NEURAL NETWORKS AND BAYESIAN DATA FUSION 64

4.1 Abstract ................................................................. 64
4.2 Introduction ............................................................ 64
4.3 Methodology ............................................................. 69
4.4 Data ........................................................................ 72
4.5 Results .................................................................... 73
4.6 Discussion ................................................................. 75

5. CONCLUSIONS ............................................................. 81

5.1 Contributions ............................................................. 81
5.2 For Further Research ................................................... 82

REFERENCES ................................................................. 83
LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Current Deep Learning Research</td>
<td>28</td>
</tr>
<tr>
<td>3.1</td>
<td>Activation function equations</td>
<td>36</td>
</tr>
<tr>
<td>3.2</td>
<td>Testing accuracy with various configurations</td>
<td>54</td>
</tr>
<tr>
<td>4.1</td>
<td>Relevant values for raw softmax data.</td>
<td>74</td>
</tr>
<tr>
<td>4.2</td>
<td>Relevant values for fused softmax data.</td>
<td>74</td>
</tr>
</tbody>
</table>
## LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>Histogram of BI-RADS rating of cases within DDSM</td>
<td>39</td>
</tr>
<tr>
<td>3.2</td>
<td>Histogram of patient age of cases within DDSM</td>
<td>40</td>
</tr>
<tr>
<td>3.3</td>
<td>Cropped image extracted from MLO view of cancerous mass.</td>
<td>41</td>
</tr>
<tr>
<td>3.4</td>
<td>Cropped image extracted from CC view of cancerous mass.</td>
<td>42</td>
</tr>
<tr>
<td>3.5</td>
<td>Color scaled down-sampled image</td>
<td>45</td>
</tr>
<tr>
<td>3.6</td>
<td>Color scaled Gaussian filtered image</td>
<td>47</td>
</tr>
<tr>
<td>3.7</td>
<td>Color scaled high contrast image</td>
<td>48</td>
</tr>
<tr>
<td>3.8</td>
<td>All three filtering steps stacked to form an RGB image.</td>
<td>50</td>
</tr>
<tr>
<td>3.9</td>
<td>All four stacked images for network input</td>
<td>51</td>
</tr>
<tr>
<td>3.10</td>
<td>Proposed Network Configuration.</td>
<td>52</td>
</tr>
<tr>
<td>3.11</td>
<td>Moving average plot across the results from the testing data arranged by</td>
<td>61</td>
</tr>
<tr>
<td></td>
<td>BI-RADS where blue is correct classification and red is incorrect using:</td>
<td></td>
</tr>
<tr>
<td></td>
<td>a) Gaussian filtering, b) original gray-scale, c) non-linear stretching, d)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>all three images stacked together.</td>
<td></td>
</tr>
<tr>
<td>3.12</td>
<td>Moving average plot across the results from the testing data arranged by</td>
<td>62</td>
</tr>
<tr>
<td></td>
<td>BI-RADS where blue is correct classification and red is incorrect using:</td>
<td></td>
</tr>
<tr>
<td></td>
<td>a) SeLU activation, b) batch normalization with ELU activation, c) batch</td>
<td></td>
</tr>
<tr>
<td></td>
<td>normalization with ReLU activation, d) batch normalization with ReLU and</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SeLU activation.</td>
<td></td>
</tr>
<tr>
<td>3.13</td>
<td>Plot of training loss of the network across 1000 epochs.</td>
<td>62</td>
</tr>
<tr>
<td>3.14</td>
<td>Plot of training accuracy of the network across 1000 epochs.</td>
<td>63</td>
</tr>
</tbody>
</table>
4.1 Plot of the loss reducing with respect to the number of epochs.

4.2 LEFT: Raw mammogram of left breast (CC view) from cancer case 1132 from the DDSM. RIGHT: The same image with the mass area circumscribed.

4.3 Histogram of BI-RAD ratings for DDSM cases.

4.4 LEFT: Cropped image taken from the case in 4.4, CENTER AND RIGHT: Two examples of training images after the augmentation step in the network.

4.5 Softmax output from CNN, each axis represents the networks determination of likeness associated with each class.

4.6 Accuracy as plotted as a function of the scaling cost. The unfused results are colored in blue were the fused results are colored in red. The x-axis is logarithmically scaled.
CHAPTER 1
INTRODUCTION

The goal of this thesis is to facilitate understanding and perhaps provide new insight into the impact and scope of breast cancer. The prognosis of those diagnosed with breast cancer is directly related to the stage at which the cancer is found, the earlier the better. Hopefully this work can provide some assistance to medical professionals by providing an unbiased perspective applying deep learning to breast cancer detection in digital mammography.

1.1 Breast Cancer Impact

In the United States, breast cancer is the dominant source of early death for women under 75 years old [44]. Breast cancer affects approximately 15% people worldwide and 12% here in the United States [44]. The current mortality rate for breast cancer is 17%. Compare the overall survivability rate with 61% in the final stages and 1% in the initial stages. It is is critical to screen women regularly to catch potential cancer as close to onset as possible [44].

The demographics of breast cancer suggest this issue is pervasive among all races and ages [22]. Breast cancer has the potential to present in anyone, even with low risk factors. The most consistent risk factor among all demographics is age. Women are more likely to
develop breast cancer later in life. As a result the American Cancer Society suggest annual screenings for women above the age of 45 [30].

1.2 Current State of Breast Cancer Diagnosis and Prognosis

There are many ways to screen and diagnose breast cancer. The least invasive procedure is the physical exam. The physical exam can be performed by anyone following the instructional guidelines published by the American Cancer Society (ACS) or other such organizations [63]. The physical exam consists of massaging the breast tissue to check for hard spots. These hard areas likely represent a sub-dermal mass.

Imaging tools are often used to identify masses or suspicious areas within the breast. The most important imaging tools are the X-ray, ultrasound, and MRI [37]. Mammograms are the most common procedure for screening breast tissue [30]. Mammography is a diagnostic medical procedure wherein breast tissue is imaged. The X-rays penetrate the breast and interact with heterogeneously distributed sub-dermal tissue to form an image. The broad use of mammograms as a screening technique is due largely to easy access to existing X-ray technology.

Sonography (ultrasound), is another common breast imaging technique, is the application of high frequency sound waves to the skin. The sound penetrates the dermal layers and echoes off of surrounding internal structures. The pattern of the echoes and reverberations are further processed to form an image. An ultrasound is common in breast imaging because of its non-invasive nature and wide spread use for other medical applications. However, the ultrasound is typically not a primary screening tool due to poor comparative
resolution offered by the procedure which leads to a higher false positive rate [38]. However, the ultrasound can outperform the mammogram in instances where the breast tissue is abnormally dense resulting in a highly noisy mammogram [58].

Magnetic Imaging Resonance (MRI) is a more modern imaging technique but can be more expensive and less common than a conventional mammogram. As such, it is typically used to inform physicians about the extent of confirmed breast cancer instances. The MRI utilizes strong magnetic fields and radio waves to generate an interference pattern which can be interpreted as a three-dimensional representation of internal tissues and organs.[55]

The gold standard for diagnosing a mass in question is a surgical biopsy [24]. A breast biopsy is performed by a physician and consists of removing some amount of the suspicious tissue for lab analysis. There are three important types of biopsies performed: The first is the incision biopsy, where the surgeon will remove a small amount of the suspicious tissue with the intention of performing oncological assays for diagnosis. An incisional biopsy is performed when the suspicious mass is small and unlikely to be immediately life-threatening. The second biopsy type is the excisional biopsy. This is often used as a last step because of the biopsy’s invasive nature as it consists of removing the full mass without first confirming the cancer diagnosis. This usually a precursor to additional surgeries to further treat breast cancer patients. The last is the lymph node biopsy, in this procedure the surgeon will remove some tissue and fluid from the lymph nodes, located under the arm, to check if cancer cells have spread to the other organ systems outside the breast. A cancer positive lymph node biopsy often leads to systemic treatment because it shows a non-local cancer cell population [18].
Breast cancer onset and development has been quantified into four distinct categories numbering one (I) through four (IV). In stage I, the mass can be no larger than 2 cm. In stage II, the mass can measure up to 5 cm. In stage III, the mass can be larger than 5 cm or advanced symptoms are apparent. In stage IV, the mass is described as metastasized in that it has spread to surrounding organs typically through the lymph nodes. All of these stages have subdivisions within them and there is potential rare exceptions which defy normal cancer behavior, but using this system the medical community has been able to further describe patient information [62].

1.2.1 Local Treatments

Breast cancer that has been identified before stage IV can be locally treated. This is desirable considering the invasive nature of the systemic approach. There are two primary local treatments that can be remarkably effective.

The first important treatment is surgical removal of the malignant mass. The amount of breast tissue removed during surgery depends on the severity and distribution of the suspicious tissue. Often surgeons will remove a slight excess of surrounding tissue to reduce the chance of missing cancerous tissue. In the most severe cases, surgeons may need to perform a mastectomy, a removal of the entire breast, to excise all of the suspicious tissue [62].

Sometimes surgery is complemented with radiation therapy. Radiation therapy utilizes concentrated high energy electromagnetic rays that destroy nucleic DNA. Any cell damaged by the radiation is unable to maintain basic cell function. There are external radiation
treatments where the patient is irradiated from a nearby machine. More modern approaches utilize implantation of a radioactive source to perform treatment [62].

1.2.2 Systemic Treatments

In cases where the cancer has metastasized (spread to the lymph nodes or other organ systems outside the breasts) a systemic approach is required. The systemic treatments are more invasive, often less effective, and have severe side-effects. Many drug based treatments are used in conjunction with one another and referred to collectively as chemotherapy. Chemotherapy treatments introduce selectively binding compounds which can either sequester or destroy cancerous cells and ideally leaving healthy cells unharmed. In recent years, the amount of breast cancer patients needing chemotherapy has reduced due to more cases being identified at earlier stages. Hormonal therapy is also often used to activate a patients native auto-immune response. The artificial hormones are constructed to activate dormant genes in the hope that the body can aid in cancer cell destruction [56].

1.3 Mass Morphology

First, we will consider the general case of a mass regardless of malignancy to establish some baseline characteristics about how and why masses appear on a mammogram. Next, let us consider the benign case and unique features which can and cannot be observed readily from a mammogram. Lastly, the malignant case will be explored. The malignant case is probably the more diverse of the two, with a larger variance among observable features.
The definition of what constitutes a mass within breast tissue is ambiguous [49]. A mammogram is considered to have a mass if it has a concentrated local increase in density [49]. The density should be at a maximum in the center of the local concentration, decreasing radially from the center. The change in density can be roughly approximated as a change in brightness of the mammogram [66]. An idealized mass will be a perfect sphere with the highest density in the center linearly decreasing as the distance from the center increases. From an imaging stand-point, the ideal mass will appear as a circle with similar characteristics [26]. The loss of dimensionality causes a large loss in the signal to noise ratio. Three-dimensional mammograms exist but are not standard for screening purposes.

Additionally, breast masses are characterized by their shape. The shape relates to the two-dimensional rendering of a three-dimensional shape and falls into four categories. The categories for the shape of a mass are round, oval, lobular, and irregular. The round and oval shapes are self-explanatory, while the lobular shapes are more severely deformed circle/sphere. The last shape, irregular, is a catch all for any mass that does not lend itself to the preceding categories [48].

The last important characteristic of breast masses is their margins. A margin is defined by its density gradient as it passes from mass tissue to normal tissue. The average gradient width can fall into many categories and most are obvious such as well-defined, slightly obscured, obscured. The non-obvious categories are the microlobulated, ill-defined, and spiculated [10]. The microlobulated margin refers to a lumpy margin. The ill-defined margin indicates a blurry margin wherein it is difficult to tell where the mass truly ends/begins.
The last case is a spiculated margin, where the mass appears to have extending flagella which infiltrate surrounding tissues [1].

Benign masses sometimes arise as a microcalcification of tissue. These microcalcifications are random accumulation of calcium salts that occur in vivo for a variety of reasons. The accumulation of calcium distorts incoming X-rays to form bright spots on mammograms. In other cases, the benign masses are caused normal cell types following a disorganized architecture leading to mass formation but with asymptomatic effects. The benign mass shape tends to be round, ovalular, and rarely lobular. The margins are well defined, so if any mass has does not fall into this category it is likely worth further exploration [28].

Malignancy results from uncontrolled non-apoptotic cell growth from cells that no longer perform normal function. The cells spread and reproduce increasing cell-density thereby crowding out normal cell growth and function. The underlying cause of malignancy is complicated and nuanced but its modus operandi is important because it informs physicians how to look for malignant masses. Malignant masses have distinguishing characteristics. Their shape is typically irregular, or in the worst case spiculated, indicating that abnormal cells are infiltrating neighboring tissues. The margins can appear in a variety of ways due to relative tissue density and its starting location within the breast [20].

1.4 Machine Learning Overview

Model-based machine learning (ML) is a method of data analysis where the model used is updated based on incoming data features to effect future output. It is a part of several fields including digital signal processing, computer vision, remote sensing, information
compression, and computer engineering. Many ML techniques can be parsed into three main categories based on outputs: Supervised, where the user supplies the algorithm with labels. Unsupervised, where the algorithm explores data distribution and structure to group data into clusters. Semi-supervised, where a small subset of data is labeled and from the labels the algorithm is able to distinguish future data and label it accordingly to reinforce and improve its interpretation of the given labeled data. Additionally machine learning can be further distinguished based on model structure. Shallow learning (SL) is a pejorative term referring to any machine learning algorithm which only utilizes hand crafted models and data features such as a support vector machine (SVM) or even a two layer neural network. The concept of a deep network comes from hidden layers which do not directly interact with the input or the output from the model. Many shallow learning applications have shown great performance but have been largely replaced on newer learning techniques. Deep learning (DL) utilizes structures normal input and output models but with hidden layers which are shaped based on the data without user input. These hidden layers can be any size, quantity, and route information dynamically through system progression.

ML applications have exploded over the past decade. In computer vision, region-based conventional neural nets (R-CNN) have successfully generated three dimensional renderings of two dimensional images [82]. In remote sensing, ML algorithms have been employed to detect oil spills based on satellite radar images [54]. In data sciences, searching algorithms have been improved using previous search information to inform downstream suggestions [40]. In the near future, self-driving cars, which utilize many machine learning concepts simultaneously, will likely have profound impacts on civilization. In many cases,
the machine can out-perform its human counter-part by orders of magnitude on both speed and accuracy of the results.

1.5 Applying Deep Learning to Medicine

How can new ML techniques be applied to the medical sciences for the benefit of mankind? This question is answered by many engineers from all over the world with new and different applications of ML concepts. This thesis seeks to answer by providing physicians with a new tool to improve patient care. The benefits of early detection previously discussed can not be understated. Many methods for screening and detecting breast cancer currently exist; however problem remains largely unsolved. There are morphological features which are suggest the presence of cancer, but their are cases which defy these norms and can result in preventable death. DL can notice underlying data features which are unnoticeable from the human perspective. Perhaps deep learning can reduce breast cancer patient mortality by helping physicians the world over to catch cancer nearest its onset.

Breast cancer detection via mammography also presents a unique combination of open questions within the DL research community. The open questions result from frequently encountered problems from a variety of DL architectures. The most daunting concern is the lack of public mammograms for use as training data. Building a deeper networks require exponentially more training data to automatically learn data features [74]. For mammography, the data is typically limited to 10,000 images. This thesis utilizes the digital database for screening mammography (DDSM), which will be discussed further in subsequent sections [35]. Lack of training data causes incomplete learning of underlying
data structures which can produce poor performance. As previously discussed, well-known mass characteristics are not always reliable for cancer determination, therefore the DL architecture needs to be deep enough to learn nuanced information encoded within the images. The availability of training data and the network depth required to produce better performance are exactly opposite forces at play in DL mammography.

Another issue in DL mammography concerns the representation of internal breast tissue as a two dimensional image. In reality, the human breast is three dimensional containing a complex substrate with many unique tissue types with a degree of variance between person to person. The lower dimensional rendering can show the mass with many unique surrounding contextual features which can have profound impacts on any attempt at classification. To combat incomplete representations, radiologists take multiple views of a single breast which increases the chance at detecting suspicious areas. Both breasts need to be considered when classifying as well. Two breasts and multiple images of each breast present an opportunity to employ data fusion techniques to unite DL outputs.

The renaissance of ML is currently underway. New applications, algorithms, etc. are rapidly conceived and implemented everyday. Herein, the application of ML to the field of mammography is examined and new techniques are utilized to produce novel and exciting results. In the next chapter, current approaches in mammography and ML will be discussed.
1.6 Contributions

This thesis makes new contributions to the ML community, specifically in regards to mammography. Most notably, two new deep learning methods are introduced and discussed. Additionally, two secondary contributions are presented.

- The first contribution, is a technique based on Bayesian statistics to aid in decision level fusion of multiple mammograms of an individual’s breasts. In the past, others have considered these images as independent when classifying however, from a patient's perspective, the individual either does or does not have cancer. This technique shows a demonstrable increase in correct classification among the testing data.

- The second technique presented in this thesis is the combination of several cutting edge convolutional neural network techniques to improve accuracy. The data is prepossessed and reconstructed to form new images allowing for the application of augmentation techniques usually reserved for color images. Improving the data augmentation reduces model over-fitting, a critical concern with the little training data available. Lastly, this technique inputs multiple images simultaneously to combat potential mass occlusions.

- The first secondary contribution is the compilation of a table which details an up-to-date compilation of current publications in deep learning mammography. The table presented shows model layouts, inputs, and outputs while also describing their results in a helpful way. It is difficult to compare studies quantitatively since there is so many facets to mammography. Some describe mass detection, others classify malignancy, while others seek to distinguish ancillary features. This table allows for future research to be bench-marked more accurately.

- Another secondary contribution of this thesis is the online publication of the dataset used. The original images are intractably large with masses located potentially anywhere on the image. Over the development of this thesis, a dataset was generated by manually selecting the center seed point of every mass. In addition to the seed points, a bounding box with eccentricity measures are including for other machine learning application needs. By creating and making publicly available, the dataset can be used for future research so that mammography studies can accurately compare results with no concern for differing mass location problems.
CHAPTER 2
BACKGROUND

Machine learning is rapidly evolving, the same can be said about its application to digital mammography. This chapter explores basic convolutional neural network (CNN) elements. Also discussed are previous incarnations of computer-aided diagnosis tools, as well as, providing an explanation of the current state of deep learning mammography.

The recent explosion in CNN applications can be traced back to several pioneering work which propelled CNN performance to the forefront of image classification techniques. These groundbreaking works include AlexNet [53], VGG Net [71], Inception [76], and ResNet [34]. These works as well as others will be introduced and discussed.

2.1 Convolution Neural Network Building Blocks

This section seeks to explain the basics of CNNs by deconstructing the concept into its simplest elements: the convolution, activation, pooling, fully connected, and batch normalization layers are discussed.

2.1.1 Convolution

The namesake layer of the CNN is the convolution layer. The inputs to the convolution layer is either the original input image or a convolution output from a proceeding layer,
hereafter referred to as the input image. Convolution is a misnomer in that its action is a correlation operation. It consists of a kernel with a user-defined size, typically of equal width and height with a depth corresponding to the number of spectral bands or image channels. The kernel traverses the input image in a raster scan style with a preset stride (typically $1 \times 1$) where the kernel weights are multiplied by the pixel values in the current location of the kernel. The aforementioned multiplications are summed and the resulting value is output as a pixel value for that location. This process repeats until the entire image is scanned. In the cases of edges, either kernel locations outside the normal image are omitted or those locations are considered to be zero.

The purpose of the convolution layer is to learn the spatial relationship of pixels directly from the data. Within images, edges can be oriented in any direction and in many different positions. The convolutional layer can account for these transitional and rotational shifts far better than a neural network which treats individual pixels as neurons [57]. CNNs have grown and changed since their conception but this most basic layer remains largely unchanged.

2.1.2 Activation Techniques

The values from the feature map produced from the convolutional layer are subjected to a step intended to identify incidents of high correlation apart from the smaller values. It is also important to treat these values such that they are always differentiable, a requirement for back-propagation. In the cases where back-propagation values approach zero or infinity, it is referred to as vanishing or exploding gradient respectively.
Early activation steps utilized a sigmoid function to squash negative values. Unfortunately, for feed forward networks such as a CNN, the sigmoid function also leads to the homogenization of values which can lead to vanishing gradients. It is also computationally expensive to calculate a sigmoid operation compared to modern activation techniques, therefore it sees little use in CNNs today.

Currently, the most common activation function is that of the rectified linear unit (ReLU). The ReLU has several attractive qualities. The ReLU operation is simple: First, if the input value is positive, it is unchanged. If the input is negative, then the output is zero. This function is differentiable everywhere except at zero. This is computationally efficient since the gradient is always one in the positive case and zero in the negative case.

New activation techniques are emerging with promising, albeit sparse, results. Some are variants of the ReLU, such as the leaky ReLU and the PReLU [21]. The leaky ReLU assigns a slope to the negative values rather squashing to zero whereas the parametric ReLU (PReLU) does the same thing but allows the negative slope to be trained along with other variables throughout the network. The idea behind their implementation is allowing some negative values to propagate through the network can have some synergistic effects, specifically in regard to succeeding convolution layers.

Exponential activations has recently been introduced. The exponential linear unit (ELU) [21] and later the self normalizing exponential linear unit (SELU) [50] utilize an exponential function and have been shown to have properties well beyond their linear counter-parts. The exponential activators handle internal-covariate shift intrinsically, without the need for
batch normalization, described later in this section, offering a performance boost with less computational burden.

### 2.1.3 Pooling

One of the biggest problems facing CNNs is that with each successive output feature map, the amount of information produced increases. The increased amount of information does not necessarily mean an increase in relevant information. Furthermore, the spatial location of information becomes less relevant as image features traverse the network. The introduction of the pooling layer solves both of these problems.

The pooling layer traverses an input image in a fashion similar to the convolutional layer. However, its operation is different in that it reduces the number of output values by performing either an averaging or a maximum operation. The most common pooling layer is the $2 \times 2$ kernel with a $1 \times 1$ stride configuration. In this configuration an input image’s width and height are reduced by half.

### 2.1.4 Fully Connected and Softmax

Once the dimensionality of the output feature maps has been reduced by successive pooling layers, the remaining network layers are often converted into fully connected layers. Fully connected layers are equivalent to hidden layers in a multi-layer perceptron (MLP). These layers are referred to as fully connected, since each neuron is connect to every other neuron in the preceding and succeeding layers.

After the fully connected layers, there is usually a final layer which contains a neuron for each class of interest in the CNN. This layer is referred to as the softmax layer and it is
critical for successful back-propagation optimization. The softmax layer values are logarithmically transformed such that they represent a probability value for the likeness of each class. The class with the highest softmax value is chosen as the most likely representation for that class.

2.1.5 Optimization

Weights and biases are trained for each fully connected and convolutional layer using back propagation. The back-propagation algorithm is shown below but in essence it seeks to follow gradient descent to find a global solution which starts with the softmax layer. An ideal network will classify an object with a softmax value of one for the class in question and zero for all others. Any deviation from that expectation allows the network to alter its weights to reflect the expected results.

2.1.6 Generalization Techniques

There are two main generalization techniques employed by modern CNNs. The first is called dropout [74], which randomly sets neurons to zero in the forward pass and does not include them in the back-propagation step. Temporarily disabling random neurons that may be already well trained allows opportunities for other neuronal pathways to be trained as well. On the testing stage, the neurons are no longer dropped and the synergistic effects of multiple trained pathways can possibly contribute to a higher correct classification rate.

The second generalization technique is batch normalization [42]. Bath normalization seeks to give an equal chance for all feature maps across a hidden layer. The batch normalization operation subtracts the mean and divides by the standard deviation of all feature
maps contained within a layer. This operation typically incorporates the entire batch, hence its name. The output from the batch norm layer is reduced to zero mean and a standard deviation of one. This reduces the contribution any one neuron can make across several hidden layers, preventing a large positive result in a previous layers from overshadowing current output as data progresses through the network.

Batch normalization performance gain is subject to conditions. In order for batch normalization to be effective it requires a batch input size to be a representative sample of the entire dataset. If the batch size is too small to adequately represent the distribution of the feature data then the batch mean and variance can differ significantly enough from the global mean and variance that noisy artifacts can be introduced into the system [42].

2.2 Convolutional Neural Network Evolution

Over the past few years major advances in the field of CNNs have lead to many changes in network width, depth, and layer configurations. This section seeks to illuminate breakthroughs that forever changed the field of CNNs applied to image recognition.

2.2.1 AlexNet

Although CNNs have been around for 1985, their rise to prominence occurred in 2012, when Krivzhevsky et al.’s network outperformed all previous image classification techniques in the ILSVRC-2012 image classification competition [53]. This new implementation of CNN was dubbed AlexNet. AlexNet used five convolutional layers with two fully connected layers that fed into a 1,000 class softmax layer. AlexNet achieved a top-5 test error rate of 15.3%.
AlexNet utilized two relatively novel concepts to reduce over-fitting. The first is dropout which is described in the prior section. The other important contribution from AlexNet is its use of data augmentation. Data augmentation allows for the artificial expansion of training data by making slight alterations to existing training images to increase the networks ability to generalize image object features rather than unwanted background information. In AlexNet’s augmentation scheme, training images were randomly cropped as well as relative values from RGB pixels were randomly manipulated, such as hue, contrast, brightness, and saturation.

2.2.2 Inception

Google’s contribution to the CNN field comes in the form of the network called Inception, or a specific architecture called Inception. The details of the network are outlined in a paper by Szegedy et al., titled ”Going Deeper With Convolutions” [76]. The authors introduce a new layer, called a concatenation layer, which combines the outputs of several different layers with different dimensionality. The concatenation layer selectively fuses the results of convolution with different kernel sizes (3 × 3 and 5 × 5) as well as a pooling layer. Inception’s biggest contribution is displaying that not all layers and network designs are created equal. Given GoogLeNet’s results, it was able to achieve state-of-the-art results compared to other networks of similar depth and width.

2.2.3 VGG Net

The next important breakthrough using CNN investigates network configuration and architecture role in overall performance. In Simonyanet et al.’s 2015 work, they suggest
that using multiple $3 \times 3$ kernels in convolutional layers can represent both small and large features [71]. Their network design is also much deeper than other networks at the time with up to 19 weight layers, dubbed VGG Net. Their results suggest each convolutional layer output can possibly be at least a factor of two greater than the previous layer to see meaningful increase in performance.

Additionally, their work explores the effect of receptive field and how it relates to object of interest size in pixels. They conclude that for successful object recognition in the network to occur, the object must reside almost entirely within the receptive field of a single neuron in the last convolutional layer. Their work on the receptive field outlined the importance of network depth. They suggest, the deeper the network the better the performance.

### 2.2.4 Residual Networks

Microsoft’s sponsored work picks up where VGG Net ends. In He et al’s 2016 work, called ResNet, the authors explore the idea that a deeper network always better [34]? Where VGG Net’s maximum depth was 19, ResNet has a depth of 152 layers. ResNet combats vanishing/exploding gradient effects by using batch normalization coupled with residual neuron connections. Residual connections feed forward both a residual representation of the previous layer and along with a normal convolutional response. Both the convolutional and residual response are fused using a special pooling layer. The process is repeated throughout the network until the fully connected layer which proceeds as normal.
Using residual connections, ResNet was able to achieve a 6.43% error on the CIPHAR-1000 dataset.

2.2.5 Network Architecture Influence

Given the above information it is obvious how much and how quickly the “best” CNN design can change. In 2012, AlexNet blew the lid off of image classification schemes using deep stacked convolutional network. In 2013-14, VGG Net and Inception models showed the important of kernel selection and network depth. Finally, in 2016, ResNet took network depth to the extreme and showed that deeper to a point is better but there is an upper bound to network depth.

From these influences, a basic structure for applying a CNN to mammography is obvious. The network will utilize $3 \times 3$ kernels for each convolutional layer, the output of each convolutional layer should expand in factors of two. Batch normalization and non-linear activation are important features to be included. Data augmentation plays a pivotal role in mammography given that data is so scarce.

There are several network parameters referred to as hyper-parameters that need to be discovered via experimentation, since there are no known rules to dictate network layout. The number of fully connected neurons and the depth of the fully connected layers is unclear. The input size of the images depends on the availability of computational power. The effect of batch normalization and overall performance is also tied to the mini-batch size and computational power. These factors need to be weighed together to find a balance between the computational burden and available resources.
2.3 Digital Database Screening Mammography

The Digital Database for Screening Mammography (DDSM) is the largest and possibly most widely-known online database containing real mammograms and cases. It contains 2,026 cases in total. The cases are split into diagnosis: Normal, benign without callback, benign, malignant. There are 695, 141, 870, and 914 cases for each diagnosis, respectively. The DDSM was specifically compiled to create a benchmark database around CAD mammography. The cases contained with the DDSM come from Massachusetts General Hospital, Wake Forest University School of Medicine, Sacred Heart Hospital and Washington University of St. Louis School of Medicine [35].

Each case file contains several images along with auxiliary information. The most important perhaps is the label appended to the name of each case file. These are the same as the diagnosis mentioned earlier and serve as class labels in the ML application. Furthermore, each case file contains four images. These images are the cranial caudal (CC) view and the medio-lateral oblique (MLO) view for both the left and right breast of the individual described by the case information. In addition to the raw images, there are overlay files which describe a region of interest (ROI) that contains a mass of some kind. The coordinates presented on the overlay files are generated from a radiologists marking.

2.4 Data Augmentation

Over fitting often occurs due to insufficient training data or prolonged training. The advent of the data augmentation schemes dramatically altered the state of neural networks. Prior to data augmentation, training a neural network required a prohibitively large amount
of data. This section seeks to explore the artificial expansion of training data using various data augmentation techniques.

The most common and easy to implement data augmentation scheme involves translations. Given the nature of CNNs rotating, cropping, and any other spatial transformation presents an effectively new image for the network to train on. In the cases were the input image is RGB or more channels there can also be manipulations across multiple channels such as overall brightness, saturation, and hue. Ideally, all of the transformations mentioned above retain the structure of what defines its class while irrelevant information is changed. On the other hand, if the transformations destroys the class identifier in the image then the network will most likely under-perform. These transformations are often tuned by a single parameter allowing the user to dynamically increase or decrease image augmentation intensity.

Rather than using simple data transformations, generative adversarial networks (GAN) have seen a large uptake as far as data augmentation [29]. GANs operate using two concurrent CNNs, one being a discriminator and the other being a generator. The generator takes in white noise and produces an image. The job of the discriminator is to determine whether the image presented is genuine or produced from the generator network. The two networks share an objective function but with opposite signs. Once trained, the generator network will be able to produce images indistinguishable by the discriminator network. These generated images can then be used to simulated data which can be added to the training set.
2.5 Data Fusion

Data fusion techniques aim to improve decision making processes by considering multiple input sources. Given the large amount of possibilities and implementations for data fusion, this review is not exhaustive. Instead, this section will review important data fusion concepts. Categorizing fusion techniques is difficult considering they appear in many incarnations in different steps of a system. Most fusion can be categorized as data association, state estimation, and decision fusion.

Data association is used to create a more descriptive representation by associating signals from multiple sources. The signals are combined and become associated with a likely target. Data association is often useful for combining heterogeneous sensor data for possible targets.

State estimation is utilizing information from multiple time-points or sources to estimate a targets likely position. Kalman filtering is probably the most well known state estimation technique. Multiple Kalman filter techniques can be utilized together with each assuming a different constant value. The model that accounts for the targets behavior and position is dynamically selected.

The most common approach for data fusion involves taking in decision output from other sources and produce a single cohesive decision for the group. This is called decision fusion. With regard to CNNs, decision fusion is probably the most relevant and likely to help in improving performance. Many Bayesian techniques are used on the output of CNNs to fuse together multiple examples of the same class [78, 9].
2.6 Bayesian Statistics

Bayesian statistics has found application in many different decision systems. Specifically Bayes’ theorem assigns a class to an observation or group of observations based on the data values. It is sometimes referred to as Bayesian inference. It is shown in the below equation.

\[ P(H_0|y)C_{10} \leq P(H_1|y)C_{01} \] (2.1)

From 2.1, the probability values are described by the \textit{a priori} values and must be approximated. The better the approximation the better the detector. Furthermore, the \(C_{ij}\) variables represent the cost of an incorrect decision. Choosing the \(j\)th class from the data truly \(i\)th class has a user selected cost. In the cases where a correct decision is made, there is typically a zero cost value used. This form of detector comes from the multiple hypothesis testing.

2.7 Receiver Operating Characteristic

The receiver operating characteristic (ROC) curve describes the ability of the detector to distinguish between different signals accurately. In this case, the detector is the CNN classifier and the different signals are the two classes, benign and malignant. Given some sample from the CNN the ROC, often plotted as a curve, allows the user to interpret how likely that sample is truly from the class predicted by the CNN. This metric is often reduced to the area under the curve or \(A_Z\). The \(A_Z\) describes the region of the probability density function (pdf) shared between the two classes. When using \(A_Z\), The higher the \(A_Z\) value,
maximally 1.0, the better performance from the detector. Alternatively, the worst case is a $A_Z$ value of 0.5 which indicates that there is no separation between the two classes.

2.8 CAD Mammography

This section seeks to give an overview on the current trends in using CAD techniques with respect to digital mammography. It is difficult to compare most studies since there are so many niches contained within the CAD mammography umbrella. This section will begin by reviewing some shallow CAD mammography techniques followed by a large table describing the success of deep learning applied to mammography.

There are some important caveats when comparing studies. The dataset may be different, this thesis utilizes the DDSM whereas other studies used InBreast database, or one compiled from neighboring hospitals. The number of classes may be different, some studies use either malignant as benign however some other studies utilize three: normal, benign, and malignant. The algorithm presented may instead attempt localize masses rather than classify them, this is a common approach as the methods for finding masses from normal tissue can inform decisions on their classification. Its also very hard considering that some studies report the correct classification rate, or the $A_Z$.

2.8.1 Shallow

Shallow learning is a term created to describe the classification systems that do not dynamically learn new features. It appears that unsupervised approaches tend to be used to for finding masses whereas supervised techniques dominate the mass classification subfield. Clustering algorithms are a shallow learning technique that is still in widespread
Singh et al. utilized K-means and fuzzy C-means clustering on breast tissue to find masses [72]. They achieved a detection rate of 85% and were able to exploit tissue property differences. In their detection scheme they analyzed random areas of breast tissue simultaneously and use an algorithm to partition them into clusters/regions. Masses show up as smaller denser clusters.

Supervised approaches have achieved decent results as well. Ball et al. achieved a high correct classification rate using a level set method for segmentation. The segmentation generated was able to generate a large set of features [8]. All of the features were combined and a decision was made using an SVM [8]. Campani et al. implemented a SVM that learned features rather than have them hand-crafted [15]. This approach resembles the training of kernel weights in a regular CNN but instead using an SVM to actually make decisions. This study illustrates CAD mammography gradually moving toward utilizing deep learning in its schemes.

Shallow learning techniques have been shown to produce good results but are limited to the scope of its designer’s bias and knowledge gaps. Deep learning approaches have become preferred methods for mammography because of the inherent variability on the size, shape, position, and density. The variability makes crafting features for mammography difficult since there are exceptions in all trends which makes them non-linearly separable.
2.8.2 Deep Learning

In 2.1 there are recent and ground breaking studies involved in deep learning applied to mammography. After the table is presented, some excerpts will be discussed with further explanation of why this study is significant.
Table 2.1 Current Deep Learning Research

<table>
<thead>
<tr>
<th>Title</th>
<th>INPUT IMAGE</th>
<th>CNN ARCHITECTURE</th>
<th>Convolutional Layers</th>
<th>Pooling Layers</th>
<th>Order</th>
<th>FC</th>
<th>CNN Address</th>
<th>Success metrics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arevalo 2015 [4]</td>
<td>CNN for mammography mass lesion classification</td>
<td>cropped and manually selected biopsies area overlaid</td>
<td>512x51</td>
<td>16x16, 32x32, 64x64</td>
<td>4</td>
<td>C2D</td>
<td>408</td>
<td>2</td>
</tr>
<tr>
<td>Cameiro 2015 [16]</td>
<td>Automated Learning of Deep Features for Breast Mass Classification from Mammograms</td>
<td>cropped and manually selected biopsies area overlaid</td>
<td>46x46</td>
<td>16x16, 32x32, 64x64, 128x128</td>
<td>5</td>
<td>Pre-Trained</td>
<td>Octave 0.70</td>
<td>DCECR</td>
</tr>
<tr>
<td>Dhungel 2015 [23]</td>
<td>Probabilistic Visual Search for Multimodal Mammography Images using Deep Learning</td>
<td>cropped and manually selected biopsies area overlaid</td>
<td>66x66</td>
<td>16x16, 32x32, 64x64, 128x128</td>
<td>4</td>
<td>Pre-Trained</td>
<td>None 0.81</td>
<td>DCECR</td>
</tr>
<tr>
<td>Dubrovina 2016 [25]</td>
<td>Computer-aided mammography using deep learning techniques</td>
<td>cropped and manually selected biopsies area overlaid</td>
<td>66x66</td>
<td>16x16, 32x32, 64x64</td>
<td>4</td>
<td>None</td>
<td>DeepInsight and ResNet DC 0.95</td>
<td>DCECR</td>
</tr>
<tr>
<td>Ertosun 2015 [27]</td>
<td>Digital mammography classification utilizing learning from deep convolutional neural networks</td>
<td>cropped and manually selected biopsies area overlaid</td>
<td>56x56</td>
<td>USDD, AxoNet</td>
<td>2</td>
<td>Yes</td>
<td>None 0.80</td>
<td>DCECR</td>
</tr>
<tr>
<td>Huynh 2016 [41]</td>
<td>Deep Learning in Breast Mass Classification</td>
<td>cropped and manually selected biopsies area overlaid</td>
<td>56x56</td>
<td>32x32, 64x64, 128x128</td>
<td>2</td>
<td>None</td>
<td>Segmentation 0.80</td>
<td>University of Chicago</td>
</tr>
<tr>
<td>Jadoon 2017 [43]</td>
<td>Three-Stage Mammogram Classification Based on Deep Feature CNN Features</td>
<td>cropped and manually selected biopsies area overlaid</td>
<td>56x56</td>
<td>32x32, 64x64, 128x128</td>
<td>2</td>
<td>None</td>
<td>CLAS and CNN output 0.80</td>
<td>Both readers and readers from 3D-Net (CNN)</td>
</tr>
<tr>
<td>Kallenberg 2016 [46]</td>
<td>Understanding deep learning applied to breast density segmentation and mammographic risk scoring</td>
<td>Patched whole image</td>
<td>24x24</td>
<td>16x16, 32x32, 64x64</td>
<td>2</td>
<td>None</td>
<td>None 0.80</td>
<td>DCECR</td>
</tr>
<tr>
<td>Kooi 2016 [51]</td>
<td>A comparison between a deep convolutional neural network and radiologist for classifying regions of interest in mammograms</td>
<td>Patched whole image</td>
<td>24x24</td>
<td>16x16, 32x32, 64x64, 128x128</td>
<td>3</td>
<td>None</td>
<td>None 0.80</td>
<td>Kool</td>
</tr>
<tr>
<td>Kooi 2017 [52]</td>
<td>Large-scale deep learning for computer-aided detection of mammographic lesions</td>
<td>Patched whole image</td>
<td>24x24</td>
<td>16x16, 32x32, 64x64, 128x128</td>
<td>3</td>
<td>None</td>
<td>None 0.80</td>
<td>Kool</td>
</tr>
<tr>
<td>Swiderski 2017 [75]</td>
<td>Deep learning in mammography mass classification in recognition of mammograms</td>
<td>Patched whole image</td>
<td>24x24</td>
<td>16x16, 32x32, 64x64</td>
<td>3</td>
<td>None</td>
<td>None 0.80</td>
<td>Kool</td>
</tr>
<tr>
<td>Zhu 2016 [85]</td>
<td>Adversarial-Deep Structural Networks for Mammographic Mass Segmentation</td>
<td>cropped and manually selected biopsies area overlaid</td>
<td>56x56</td>
<td>32x32, 64x64, 128x128</td>
<td>3</td>
<td>None</td>
<td>None 0.80</td>
<td>Kool</td>
</tr>
</tbody>
</table>

28
2.8.2.1 Deep vs. Shallow

Probably the most straightforward study included in the list is the study done by Arevalo et al., which classified masses using a CNN [4]. This study is important because it explores the change from hand-crafted features and automatically learning new features instead. Using then current shallow techniques they were able to achieve an $A_Z$ of 0.79. Then using a CNN they were able to boost the $A_Z$ to 0.86. Its also interesting to note how small the CNN actually was. It had two convolutional layers and was able to outperform far more conventional shallow architectures.

Another study showed the performance of shallow vs. deep learning techniques is Dhugnel et al.’s 2016 paper regarding automated feature learning for a mammogram mass classification [23]. The main contribution from this study is that the author utilizes a previous method they had published against a deep learner and compared results. They were able to achieve a 5% boost to performance by using a CNN to generate learned features rather than the hand crafted features which are usually focused on geometric and textual effects.

Although even before deep learning was inducted into CAD mammography, CAD tools have shown to be more effective at screening than physicians, it is important to benchmark progress against experts in the field. To this end, Kooi et al. performed an experiment. Their work was a comparison of using a CNN to detect masses against trained radiologists [51]. Their results suggest that deep learning can outperform shallow as well as trained specialists. The author took their results to the next level when he introduced a CNN which can output a bounding box around a classified mass [52].
2.8.2.2 Unsupervised Approaches

Unsupervised approaches have also moved toward deep learning techniques. Ertosun et al. combined an unsupervised engine which utilized a R-CNN for localization of masses and a CNN for discrimination. [27] The two networks where trained using different training data. The discriminator CNN is used to determining if an image likely contains a mass, if so it is sent to the R-CNN to draw a bounding box around suspicious mass. This study is interesting because it tries to do both localization and discrimination tasks together. The authors achieved a successful mass localization $A_Z$ of 0.85, and 0.9 false positives per image.

A very interesting unsupervised application by Kallenberg et al. used an auto-encoder to produce a sparse representation of the mammogram which was then fed into a typical CNN [46]. Using their proposed scheme they were able to achieve a $A_Z$ of 0.59. While its performance is lacking it is important to note that the authors are attempting to perform localization and classification in a single step. The autoencoder served an interesting purpose, shrinking and denoising the images at the same time which allowed for easy input into a CNN.

2.8.2.3 Supervised Approaches

Several of the most successful supervised approaches utilize transfer learning to overcome the lack of available training data. Huynh et al. utilize an a pre-trained CNN and compare its results with that of an SVM [41]. They reported an $A_Z$ of 0.81 for the SVM and 0.86 for the CNN using only 216 mammogram cases. The most significant result from
this study is that the deep learner was able to outperform the hand-crafted features even with such a low amount of training data.

Mass segmentation is also an important sub-field of CAD mammography. A 2016 work from Zhu et al. shows utilizing adversarial training techniques coupled with conditional random fields can lead to very accurate mass segmentation [85]. This article has some important insights. First, they used a fully convolutional network rather than a CNN that has several fully connected layers at the end. Keeping the network fully convolutional allows for far better analysis of spatial relationships were are totally lost once a traditional CNN goes into fully connected layers. Another important step is the conditional random fields incorporation into the network increases its ability to recognize and learn super structural features.

Mass classification is another important sub-field to deep learning mammography. In 2012, Jadoon et al. trained a CNN with features extracted from an SVM [43]. In this work, the authors classified masses as either normal, benign, or malignant. This study shows how easily CNNs can be incorporated. The CNNs input is a tiny $28 \times 28$ image with features extracted using the discrete wavelet transform and contrast limited adaptive histogram equalization (CLAHE) pre-processing steps.

The current best performing CAD mammography classification scheme is described in Carneiro et al.’s 2015 work. They take two pre-trained CNNs and further train them on MLO and CC views independently [16]. They achieved a $0.90 A_Z$ on the DDSM as well as $0.90$ on the InBreast dataset.
CHAPTER 3

UNREGISTERED MULTI-VIEW MAMMOGRAPHIC IMAGE CLASSIFICATION
USING SELF-NORMALIZING CONVOLUTIONAL NEURAL NETWORKS

3.1 Abstract

This work seeks to combine image processing, data augmentation, self-normalizing neural networks (SNNs) and apply them to existing public mammographic data. The digital database for screening mammography (DDSM) was utilized for the purposes of this work. Each case within the DDSM was significantly reduced in size using the Haar discrete wavelet transform. The reduced pictures were subjected to three different image processing schemes and were then stacked to create faux RGB images. Complementary views were input into the network simultaneously to allow for richer context of the region of interest (ROI). The results from the study were manipulated to compare system performance and are presented herein. The model was able to achieve competitive results with a ten-fold cross validated testing accuracy of 76.9% by using batch normalization with ReLU activation on convolution layers and scaled exponential linear units (SELU) on fully connected layers. The effect of adding age to improve model performance is also investigated and discussed.
3.2 Introduction

3.2.1 Digital Mammography

Prognosis of breast cancer patients is directly related to the interval between the cancers onset and its discovery. [44] With the advent of digital mammography, physicians have been able to detect and localize breast cancer early in its development [63]. In recent years, machine learning has seen widespread application to the medical imaging community. Specifically, deep learning seems uniquely appropriate given the high variance of symptomatic and morphological characteristics between individuals. Deep learning can key in on non-obvious features without the input from the user.

Masses found within the breast have been categorized to standardize diagnosis [67]. These metrics include quantitative features such as size and location, along with quantitative descriptors such as shape and margin. The categories for shape include round, lobulated, ovoidal, and irregular. For margin, descriptors include well-defined, ill-defined, and spiculated [24].

Some categories used to describe breast masses are diagnostically relevant to breast cancer. Masses that appear as round, or ovoidal, with well defined margins are more likely to be benign and not require further treatment [20]. Additionally, masses with irregular shape and ill-defined margins are more likely to be cancerous [18]. A mass with obvious spiculation, mass surface appears to have spikes or points along its surface, is the most consistent indication of breast cancer [20].

Although masses are three-dimensional in reality, they are rendered as two-dimensional figures when viewed on a mammogram. No single view of an individual breast is sufficient
for screening purposes. Physicians will often image the breast from multiple angles to reduce the possibility of a false negative. Most commonly two vantage points of a breast are imaged during a screening, the medio-lateral oblique (MLO) view and the cranio-caudal view. The MLO view is taken from in-between the breasts at a raised angle where the CC view is taken from the top downward. Breast density varies from individuals and micro-calcifications, small calcium deposits within breast tissue, and random tissue imperfection makes definitive classification of a mass as either benign or malignant very difficult [28].

The shortcomings of mammograms are supplemented by magnetic resonance imaging (MRI) as well as 3-D X-rays to provide new information about suspicious areas within breast tissue. A mass can be most accurately classified by surgical biopsy [37]. Mammograms are still regularly used as a screening tool because other procedures are more expensive, invasive, or inaccessible for many. [55]

3.2.2 Deep Learning Mammography

Applying deep learners to solve real world problems has been the driving force in development in progress in the field, and medical science has benefited from this journey. Specifically, mammography has seen a multitude of unique applications of deep learning with differing degrees of success. It is difficult to quantitatively compare studies since the publicly available data is sparse, as well as differences in systemic goals.

In Carneiro et al.’s 2015 work, the authors detail a network architecture wherein multiple views of a ROI are input into a CNN simultaneously [16]. They were able to show significant performance gains without the need for image co-registration. Co-registering
images involves determining images relative positions to one another to generate a 3D coor-
dinate system shared between the images. Their model performed at 0.90 ROC predicting BiRADS for masses. Their experiments utilized the DDSM and InBreast as testing data. The BiRADS rating system will be discussed further in the next section but this is distinct from classifying a mass as benign or malignant. The same author expanded this model to create a fully automated system which while it did not offer significant performance gains it did offer a greater utility by removing the need for user input [17].

Another study utilized a similar multiview architecture wherein the raw images where down-sampled prior to network input to reduce the noise present as well as reduce the computational burden for a single image input, thereby increasing the batch size. [59] In addition many deep learning mammography studies have included ancillary data to improve network performance. Most notably the effect of age [77] as well as BIRADS rating [13] have been shown to have a profound impact [20].

There have also been some holistic studies on the relative performance of CAD based mammography, showing mammography’s increasing dependence on deep learning approaches [64]. Additionally, another work investigates which cases are difficult for different designs [12].

Currently CAD mammography is dominated by deep learning approaches. Some of the more successful strategies involve using transfer learning to overcome the training data deficient. [41] Others complement transfer learning with severe data augmentation steps to artificially expand the training data [85]. Some of the more successful approaches have
employed shallow learning techniques such as an SVM coupled with a CNN to reduce CAD system dependency on large training data availability [27].

### 3.2.3 Self Normalizing Neural Networks

The advent of SNNs pumps new life into a relatively old technique. Feed-forward neural network applications have steadily been replaced by newer network architectures. Sequenced based applications now use recurrent neural networks (RNN) such as long-short term memory (LSTMs). Image based approaches use CNNs or a combination of RNN and CNN construction. This replacement is largely due to the exploding/vanishing gradient problem that is the limiting factor in increasing network depth [50].

In Table 3.2.3, the equations for common activation functions as well as the newer scaled exponential linear unit (SELU) is presented. Notice that the SELU is a constant scaled version of the ELU activation. These parameters have shown to accomplish batch normalization without the need for increased computational burden on the graph. This could result in a huge amount of savings for very deep networks.

<table>
<thead>
<tr>
<th>Activation</th>
<th>Equation</th>
</tr>
</thead>
<tbody>
<tr>
<td>ReLU</td>
<td>( f(x) = \begin{cases} 0 &amp; x \leq 0 \ x &amp; x &gt; 0 \end{cases} )</td>
</tr>
<tr>
<td>ELU</td>
<td>( f(x) = \begin{cases} e^x - 1 &amp; x \leq 0 \ x &amp; x &gt; 0 \end{cases} )</td>
</tr>
<tr>
<td>SeLU</td>
<td>( f(x) = \lambda \begin{cases} \alpha(e^x - 1) &amp; x \leq 0 \ x &amp; x &gt; 0 \end{cases} ) where ( \lambda = 1.05 ) and ( \alpha = 1.67 )</td>
</tr>
</tbody>
</table>
In order to use the SELU activation, the authors show that a unique initialization is required such that the weights can asymptotically approach zero mean and unit variance. Unfortunately the authors do not describe an initialization for convolution layers which may cause poor performance. Furthermore, the authors describe a new dropout layer, called alpha dropout, which requires a significantly reduced drop out rate of 0.1 compared to the more traditional value of 0.5. They claim that the model is still able to generalize well but any dropped values are compensated by weighting the kept values by $\left(\sqrt{n^{-1}}\right)$ where $n$ is the number of neurons dropped in any given layer. The reduction in the dropout value also has benefits of earlier model convergence leading to a significantly lower training time.

### 3.3 Data

Overall CNN performance is reliant on large amounts of data for training. For mammography, there are databases available for CAD purposes, yet these resources are often insufficient for deep learning applications. This work utilizes the largest publicly available database for screening mammography. Other studies have included multiple mammography databases such as InBreast or regional hospital resources but this work seeks to use the DDSM as a benchmark for comparison of future works.

#### 3.3.1 Digital Database for Screening Mammography

The DDSM is an invaluable resource for deep learning mammography applications. Its was compiled by a collaboration of several universities, hospitals, and cancer research centers. Principle funding was supplied by the DOD with the expressed purpose of promoting image analysis by providing researchers with public resources [35]. The DDSM
contains 2,650 cases, with patient distribution covering a wide variety of demographics but is primarily white, 56%, and all female. Ground truth for cases was established by qualified radiologists. Cases are labeled as: benign without callbacks, benign, cancer, and normal. In addition to classification, the radiologist also provides an encoded mask which circumscribes the region of interest (ROI) for images containing a mass.

A single case within the DDSM contains one “ics” file, one MLO and CC mammographic image for each breast, and an “OVERLAY” file for cases containing an interesting region. The “ics” file lists some patient information along with logistical information about when/how the scan was performed. The ”OVERLAY” files contain an encoded mass along with a quantitative and qualitative information about the mass. There may be information about more than one mass in a single “OVERLAY” file with ground truth for each individual mass.

Among the mass descriptions in the “OVERLAY” file, there is an important descriptor called the Breast Imaging Reporting and Data System (BIRADS) rating. BIRADS ratings range from 1-5, with 1 being the least threatening and 5 being highly suspicious of malignancy. Cases with a BIRADS rating of 0 means an incomplete evaluation. A histogram of the cases utilized in this work are shown in Fig. 3.3.1. The DDSM cases have a skewed distribution of the histogram towards a high BIRADS rating.
Another important ancillary descriptor included with each DDSM case is the patient age. In Fig. 3.3.1, the distributions of both benign and cancer cases are shown in a bar graph with a bin size of 10 years. The cancer cases have an average age of 58 and a standard deviation of 5 years. The benign cases have an average age of 52 with a standard deviation of 3 years. From this figure it is apparent that age has a significant influence on the presence of cancer in an individual.
In addition to deep learning mammography, the DDSM has seen applications in many mammography CAD systems. Breast density is an important metric in cancer risk assessments. In Bosch et. al.’s work, the authors use the DDSM to quantify the breast tissue density of each mammogram using a support vector machine classifier [13]. Breast density classification has also been used to predict the BIRADS value using statistical information extracted from the raw mammogram images [19]. Other works have utilized unsupervised techniques to analyze mammograms within the DDSM. One such work is from Verma et al.’s 2011 work wherein features are extracted from the mammograms, the features are clustered. The resulting clusters are fused to generate classifications [80]. Another unsupervised approach utilized the DDSM to provide unsupervised segmentation of masses. The segmentation output was further analyzed to extract features that would normally be difficult to determine such as mass morphological features [32].
3.3.2 Seed Point Acquisition

The DDSM requires significant modifications before it can be successfully utilized. In order to homogenize mass location within the images extracted from this work, a seed point for each mass was needed. There have been many attempts to automate seed point selection for many reasons although typical frameworks select seed points with mass segmentation in mind [84, 68, 11]. For this work, the seed point was manually selected by applying the mask generated by the overlay file to the raw mammogram image. Within a $9 \times 9$ pixel area of the manually selected point, a max search is operating is performed to change the seed point location to the maximum value of the pixel neighborhood. This allows for less variance when applying the Gaussian filtering described in the image processing steps below. The MLO view of a cancerous mass after the seed point is selected is shown in Fig. 3.3.2.

![Figure 3.3 Cropped image extracted from MLO view of cancerous mass.](image)
3.3.3 Pre-processing Steps

Once a seed point is selected for each image there are several important steps to prepare the extracted image for further downstream processing. Images were extracted from the raw mammograms by cropping a $2048 \times 2048$ region as well as a $1024 \times 1024$. Many seed points were found very near image borders. In these cases, a full image was extracted by zero padding any non-existing pixels. The padded images were then manually referenced between multiple views such that each case mass has an MLO and corresponding CC view associated with it. The image shown in Fig. 3.3.2 has a companion view which required padding is shown in Fig. 3.3.3.

Figure 3.4 Cropped image extracted from CC view of cancerous mass
3.3.4 Augmentation

This work requires a heavy amount of augmentation techniques to artificially expand the database due to insufficient training data availability. The augmentation techniques used include random flipping, cropping, rotating, and brightness adjustments. These steps are outlined by the famous AlexNet paper which suggest that small pixel adjustments create entirely new images for the network to learn on even if the focus of the image remains unchanged [53]. Previous experiments by the authors of this work used the typical gray-scale mammogram pixel values as network inputs and determined by experimentation that the maximum value without yielding diminishing returns was a maximum 10% brightness adjustment.

Augmentation on the gray-scale images alone was insufficient in preventing network over-fitting. This work utilizes several image processing techniques to generate new images which are stacked on top of one another to generate new faux color (RGB) images. This allows for the implementation of augmentation steps described by Krizhevsky et. al [53] wherein the saturation, hue, and contrast can also be randomly changed to prevent early model over-fitting. Through experimentation, the maximum value without yielding diminishing returns was a maximum 2% adjustment to these operations.

Without RBG manipulation the apparent database to the network is about 1.4 million images. That number jumps several orders of magnitude upon introduction of the color augmentation steps. These augmentation steps occur in a randomized order as well as randomized in their respective parameters bounded by the values mentioned above.
3.3.5 Down-sampling

Previous incarnations of the work presented here attempted to use the full sized image as network input. Unfortunately, when using the full sized image led to network non-convergence problems producing poor results with a testing accuracy of 67% whereas images which where down-sampled by a factor of 8 produced slightly better results of 68%. Therefore, an effective system of down-sampling was introduced to improve network convergence. The down-sampling used in this work closely resembles the technique described by Hilton et al. [36]. The Harr wavelet transform was used to down-sample the $1024 \times 1024$ and $2048 \times 2048$ raw images to a size of $512 \times 512$. The images produced using the $1024 \times 1024$ raw images are hereafter referred to as zoomed in while the images produced using the $2048 \times 2048$ raw images are referred to as zoomed out. The zoomed in image generated from the raw CC image in Fig. 3.3.3 is shown in Fig. 3.3.5 as a color scaled image.
3.3.6 Gaussian Filtering

Gaussian 2D filtering has long been used as a way to enhance images by reducing the background pixel values. Commonly, this is used in for segmentation processing where the seed point is the center of a growing circle and steep drop offs in pixel values can be used as edge detection indicators [14, 31]. The Gaussian filter presented in this work is based on Ball et al’s 2007 work. [8] The filter presented in Alg. 1 is identical to their work except that the 0.2 coefficient used in step 4 was originally 0.4. This adjustment was made based on experimental findings and is required due to the discrepancy in image sizes/resolution.
Algorithm 1: Gaussian filter algorithm

Result: $E$ - Gaussian filtered image

1. $I$ - mammogram image, seed point as inputs;
2. $P$ ← convert $I$ to polar coordinates with center at seed point;
3. $\mu(r) ← \text{mean}\{P(r, \theta)\}$ for all $r$; $\mu_{\text{max}} ← \text{max}(\mu)$;
4. $\sigma$ ← smallest radius $r$ with $\mu(r) \geq 0.2\mu_{\text{max}}$;
5. $G(r, \theta) ← \begin{cases} 1 & r \leq \sigma \\ \exp\left(-\frac{1}{2}\left[\frac{r-\sigma}{\sigma}\right]^2\right) & r > \sigma \end{cases}$;
6. $M$ ← convert $P$ back to Cartesian coordinates;
7. $E ← M \times I$ element wise multiplication;
8. $E ← (E - \text{min}(E)) / \text{max}(E) - \text{min}(E)$;

The Gaussian filter described should provide a unique perspective from other techniques presented in this work since it is the only processing step that is radially based. The filter works by converting the image into a polar space. The mean of the values at a radius (at any angle) is calculated and compared to the seed points value. Once the value has decreased by a relative amount, the Gaussian filter begins to squash pixel values with radii outside the boundary. An example of this filters effect is shown in Fig. 8 that used Fig. 3.3.5 as input.
Notice from the original image in Fig. 8 and the Gaussian filtered image in Fig. 8, that the mass has clearly defined edges after filtering. These results are not necessarily typical because in some cases the mass is in the background of the image which can result in poor filter output.

3.3.7 Non-linear Stretching

Applying non-linear stretching, or stretching acts as contrast enhancement. Many techniques for contrast enhancement techniques exist with most image based applications focused on reducing ambient noise will preserving edges and underlying structures [83]. New contrast enhancing techniques resemble histogram equalization operations where local statistics are used as input parameters for contrasting [5]. The algorithm presented in
Algorithm 2 is based on typical contrast framework but also includes a preceding denoising step using the a bounded sigmoid function.

**Algorithm 2:** Non-linear stretching algorithm

**Result:** $S$ - stretched image

1. $I$ - mammogram image, seed point as inputs;
2. $\mu \leftarrow$ mean of all pixels in $I$;
3. $T \leftarrow \frac{1}{1+\exp(-0.25\mu(I-0.75\mu))}$ for all pixels;
4. $T \leftarrow T^2$ for all pixels;
5. $S \leftarrow T \cdot I$ element wise multiplication;
6. $S \leftarrow (S - \min(S)) / \max(S) - \min(S)$;

The variations in breast density between images makes it difficult to generalize parameters for reliable effect across all images. Therefore the algorithm in Alg. 2 utilizes images statistics to squash potential noise values using a sigmoid function. Using Fig. 3.3.5 as input, the output image from nonlinear stretching step is shown in Fig. 6.

![Figure 3.7 Color scaled high contrast image.](image-url)
The noise within the mammograms are both a result of the equipment used to produce the image as well as internal variations in breast tissue. The intent of the contrast enhancement described is to weight pixel values by their relative distance to the camera to produce some semblance of a background vs. foreground regardless of the seed point selected. This is shown in Fig. 6, consider the changes in brightness that can inform a model about pixel depth and intensity.

3.3.8 Data Input Structure

After completion of the image pre-processing steps, the network inputs were arranged so that two views of each mass were input along with two resolutions of those views, totalling in four images input into the network at a time. Prior to network input, the final step in pre-processing for single images is to stack them in a way that they form a color image. The color image generated from the images in figures 3.3.5, 8, and 6. It important to note that those figures are color mapped and are actually gray-scale. The resulting image is shown in Fig. 3.3.8. The red channel in the image is the output from the nonlinear stretching step. The output from the Gaussian filtering step is affixed to the blue channel. Lastly, the green channel is the original down-sampled image.

A sample of all four images prior to network input is shown in Fig. 3.3.8. Notice that the top left is the same image from Fig. 3.3.8 while the bottom left is a zoomed out representation of the same image. The same can be said for the right-side for the close up and zoomed out representations of the same image. In some of the experiments performed
in the case of this study the age of patients was included with network input. The age value bypassed the convolutional layers and was first input at the first fully connected layer.

Input images were also subjected to image normalization to further homogenize input data. Any case that did not have multiple views or age data available were not used to train the network. Also omitted are any instances where the seed point selected is within 25 10% of picture size

3.4 Methodology

This work utilizes many well known CNN elements while trying alternate arrangements and hyper-parameter values to measure performance changes. Some well recognized elements include convolutional layers with $3 \times 3$ kernel sizes immediately followed by an activation and subsequent convolution layer. After the second convolutional layer
another activation layer is applied and finally a max pooling layer with a $2 \times 2$ stride reduces the feature maps length by 2. This convolutional block is based on the well known VGGnet’s configuration [71].

Every experiment described trained for 1000 epochs with a training rate of $10^{-4}$ using an AdAM optimizer. In the cases where batch normalization was used, the batch normalization layer was inserted immediately behind all convolutional layers. Once the size of the feature maps from the final convolution layer was reduced to $8 \times 8$ the network was fully connected. All fully connected layers included a dropout layer, which could be either a conventional dropout layer or an alpha dropout layer when used with SELU activation. The keep probability for the dropout and alpha dropout layers was 0.5 and 0.9 respectively.

During training all neurons where initialized using a truncated normal distribution, in the cases where SELU activation was used the initialization was changed for any layer
directly connected to the activator in accordance with the authors who proposed its use. 
Training accuracy and training loss were measured and recorded during training.

3.4.1 Neural Network Configuration

The network architecture has two large sections. The first network section is four parallel convolutional neural networks using the four input images described earlier. There are a total of ten convolution layers per parallel CNN. The output feature maps for the convolution layers in descending input image size is: 32, 32, 32, 64, 64, 64, 128, 128, 256, and 256.

The second section of the network is fully connected with 5 layers with adjoining dropout and activation. The number of neurons in these layers, 2048, 2048, 2048, 1024, 512. In the cases where, age was included on input data the first fully connected layer is increased by one in size. The final fully connected layer is connected to the soft-max regression layer.

Figure 3.10 Proposed Network Configuration.
The overall network construction is represented in Fig. 3.4.1 with each of the four parallel CNNs being input into the first fully connected layer.

### 3.4.2 Experiment Schedule

Many experiments were performed over the course of this work by instantiating small changes in input shape or network hyper-parameters. The first round of experiments involve testing the performance of all the pre-processing steps individually and then testing the effect of the combination of the three into the faux RGB images. When only using a gray-scale image, the data augmentation scheme is reduced to operations that can be performed on single channel images.

The second round of experiments explores the performance change when using different activation functions. When using ReLU and ELU activation, batch normalization is included in the network. Furthermore, a hybrid approach is applied wherein the convolutional layers use batch normalization with ReLU activation while the fully connected layers apply a SELU activation. The final experiment explores the effect of age on network performance. For this experiment hybrid convolutional ReLU and fully connected SELU (ReLU-SELU) is utilized. As mentioned before, this alters the size of the first fully connected layer by one.

In addition, the best performing configuration, in this case the stacked images with age using ReLU-SELU activation will be subjected to a ten-fold cross validation to gauge generalization properties of the network. The experiments described in this work were performed using Google Cloud Platform’s virtual machine (VM). The VM’s had four NVIDIA
Tesla P100 GPUs. Tensorflow version 1.4.1 was as medium for network generation and GPU utilization.

3.5 Results

The results generated by this work are outlined in Table 3.2. The overall best performing configuration was observed to be stacked images with age using ReLU-SELU activation to achieve a 77.2% testing accuracy. The worst performing configuration used gray-scale Gaussian filtered images achieve a 70.1% testing accuracy. A ten-fold cross validation yielded an average result of 76.9%. When age is added to the input pipeline, the model performance increased by 3%.

<table>
<thead>
<tr>
<th>Activation</th>
<th>Batch Norm</th>
<th>Image Input</th>
<th>Age Input</th>
<th>Testing Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ReLU - SELU</td>
<td>X</td>
<td>RGB</td>
<td>X</td>
<td>77.2</td>
</tr>
<tr>
<td>ELU</td>
<td>X</td>
<td>RGB</td>
<td>X</td>
<td>75.6</td>
</tr>
<tr>
<td>ReLU</td>
<td>X</td>
<td>RGB</td>
<td>X</td>
<td>75.6</td>
</tr>
<tr>
<td>SELU</td>
<td></td>
<td>RGB</td>
<td>X</td>
<td>73.0</td>
</tr>
<tr>
<td>ReLU - SELU</td>
<td>X</td>
<td>gray-scale using Gaussian filtering</td>
<td>X</td>
<td>70.4</td>
</tr>
<tr>
<td>ReLU - SELU</td>
<td>X</td>
<td>gray-scale using original values</td>
<td>X</td>
<td>74.6</td>
</tr>
<tr>
<td>ReLU - SELU</td>
<td>X</td>
<td>gray-scale using non-linear stretching</td>
<td>X</td>
<td>72.6</td>
</tr>
</tbody>
</table>

3.5.1 Channel Manipulation

Each pre-processing technique described produced unique performance across the testing values. The best performing gray-scale input configuration used the original wavelet down-sampled image at a testing accuracy of 74.6%. As mentioned, the worst performing
gray-scale input configuration used the original wavelet down-sampled image at a testing accuracy of 70.1%.

Testing accuracy alone is limiting when considering network performance. To further compare model performance Fig. 3.5.1 was generated from the softmax outputs from the trained network. The x-axis on Fig. 3.5.1 is split into two sections, the first benign and the second cancerous separated by the horizontal dark line on the plot. Within the separation for each class, each case is arranged in ascending order according to BIRADS rating. The order testing cases is constant across all subplots within Fig. 3.5.1. The colored response shown on the y-axis is a moving average representation of percent correctness. Correctness in this instance refers to whether the softmax probability from the network matches with the ground truth classification. For example, if a case is classified as 100% likely to be benign when it is actually cancerous, the figure will represent this mistake as a negative one and shaded red. The moving average allows for the visualization of local response.

There are some interesting trends within Fig. 3.5.1. Notice that using Gaussian filtering produced a highly variant softmax response. All three separate gray-scale models have similar response patterns with peaks and valleys sometimes consistent between all four subplots. When the response patters are inconsistent the response for the stacked images tends toward zero.

### 3.5.2 Activation

Different activation layers had a similar distribution of testing accuracy compared to the image processing comparison discussed above. The best performing activation function
observed is hybrid ReLU-SELU activation which achieved a testing accuracy of 77.2%. The worst performing activation function investigated is the SELU activation alone resulted in a testing accuracy of 70.1%. Similar to Fig. 3.5.1, Fig. 3.5.2 was generated to compare the correct softmax output.

The activation function used in the network plays an important role in model generalization and epochs needed for training convergence. Plots of training loss and training accuracy, shown in Fig. 3.5.2 and Fig. 3.5.2 respectively, can suggest successful convergence and early training terminating to avoid over-fitting. Notice in Fig. 3.5.2, that the training accuracy and loss converged first among the candidates. Conversely, the batch normalization with ReLU and ELU took the longest to converge however they produced better testing performance. In the middle of both the training accuracy and loss figures lies the hybrid ReLU-SELU networks, plotted in yellow.

3.6 Discussion

The proposed network configuration boasts competitive performance among other CNN based breast mass classification systems by utilizing well known convolutional construction and the newer SNN construction for the fully connected layers. Using different gray-scale image processing types separately performed worse than their combined performance. Age increased performance slightly. It is important to keep in mind that good neural network architecture is anecdotal at best and depends heavily on the dataset utilized. Therefore the scheme proposed in this work is likely not generalizable to datasets unrelated to medical classification tasks.
3.6.1 Channel Manipulation

The response patterns from Fig. 3.5.1 suggest that there is synergy between using all three image processing techniques together. The Gaussian filtering response is by far the most distinct with having several areas where the average correct response is negative. All three of the gray-scale networks’ response shows many areas where the network is very certain about each case individually whether or not the prediction is accurate. This could suggest that the network has over-fit on the training data. On the other hand, the combined color image in 3.5.1 part d shows the network is more often correct but is on average less certain about cases. This shows a successful artificial expansion of the dataset using color based data augmentation scheme.

This work used three different image processing techniques to combine these images into an RGB image to be sent into the network. These steps were selected based on the previous successful applications from the authors. This is not to say that these results will be typical nor will they always produce better results rather this work is meant to suggest that training data expansion is possible and potentially useful by following a system similar to the one described here. It is also important to note that there is no reason to stop at three channels. Three was utilized as a proof-of-concept here but the number of channels could be raised to any number.

Beyond increasing the amount of data augmentation operations available, the purpose of the image processing steps allows the user to tailor the network to target specific relevant information within the images. The Gaussian filtering is perhaps the best example since it is reliant on a good seed point and that the mass in question needs to be circular to
generate useful output. In the cases where the mass fails to meet this criteria or is otherwise obstructed or obscured in the image, the Gaussian filter will produce lackluster images for network input. The image processing steps selected for this work allowed for the input of author bias as far as what seemed to be relevant information for the classification scheme as a whole.

Finally, the image processing performed herein could have systemic effects coupled with the activation used. It is difficult to fully quantify the effect on model performance one has on the other and so for the purposes of this work, both the network activation function used as well as the image processing scheme used where held constant with regard to one another.

3.6.2 Activation

The choice in neuron activation played a significant role in network performance. From Table 3.2 it is obvious that the hybrid ReLU-SELU approach out performs any other scheme tested. It is interesting however that the SELU only without batch-normalization performed the worst. As a further experiment, a SELU only network with batch-normalization was constructed and produced slightly worse results and so was not included in the table. The purpose of batch-norm layers are to change all outgoing data in a layer to be unit-norm and zero mean, the SELU activation does this automatically without increasing the number of variables to the graph. A completely SELU based CNN is attractive because of the reduction in variables required and enforcing unit norm and zero mean to data can introduce noise artifacts especially when using smaller batch sizes.
There is very little research currently available using SELU activation in CNNs. It seems as though the pooling layer is potentially at fault for poor network performance. Similar issues have occurred when using generative adversarial networks where the convolution 2-D transpose cannot properly function when preceded by a pool layer. Therefore it is likely that a new network element will need to be developed for successful SELU application to convolution layers.

Consider figures 3.5.2 and 3.5.2, and notice that in both cases the training accuracy as well as the training loss converge the quickest when using the SELU only network. This trend is likely due to the limited dropout rate suggested when using alpha dropout available. From these figures the alpha dropout does seem to enhance convergence however it also seems to lead to over-fitting. In the cases where the hybrid activation was utilized it seems to have found a middle ground between conventional activation as well as the SELU only activation.

The response patterns shown in Fig. 3.5.2 are unique to every activation technique. Interestingly, the SELU only activation has the least amount of negative average segments in the graph but note that the hybrid ReLU-SELU network seems to have large areas of correct predictions among its peaks. From the figure it appears that both ReLU and ELU alone have large areas of poor performance. It is unclear what generated these artifacts.

3.7 Contributions and Future Work

This work produced some exciting results for deep learning mammography. Most importantly, network performance was increased when using multiple image processing step
on data input in parallel. The output from those steps was stacked on top of one another and improved testing accuracy by 2.8%. This allows for user bias input as well as increased augmentation operations available to the user.

This work also showed that including ancillary data into the network at run-time also increased performance by a similar amount, 3.2%. We propose and test a new form of convolution setup wherein the convolution layers of the CNN are constructed with typical elements, batch-normalization, ReLU activation, and max pooling, while the fully connected layers are changed to use SELU activation. In addition, the fully connected layers use alpha dropout rather than the conventional dropout. This configuration out performs any other tested here.

In the future, this work will be expanded to include the image processing steps inside the network. On input, the processing operations can be parametized allowing the network to learn optimal image processing steps. These processed images can be stacked by the network to form the faux RGB images described or they can be pushed into separate networks. If they are pushed into different networks a decision level fusion mechanic will be introduced and investigated.

Other possibilities for future research include a thorough investigation of initializations as well as new pooling techniques. The potential of SELU activation is limited currently by its lack of adaptation to the CNN scheme. SELU convolution is attractive due to its Independence of batch size which is significant for larger image CNN design.
Figure 3.11 Moving average plot across the results from the testing data arranged by BI-RADS where blue is correct classification and red is incorrect using: a) Gaussian filtering, b) original gray-scale, c) non-linear stretching, d) all three images stacked together.
Figure 3.12 Moving average plot across the results from the testing data arranged by BI-RADS where blue is correct classification and red is incorrect using: a) SeLU activation, b) batch normalization with ELU activation, c) batch normalization with ReLU activation, d) batch normalization with ReLU and SeLU activation.

Figure 3.13 Plot of training loss of the network across 1000 epochs.
Figure 3.14 Plot of training accuracy of the network across 1000 epochs.
CHAPTER 4
RAPID SCREENING MAMMOGRAPHIC ANALYSIS USING CONVOLUTIONAL NEURAL NETWORKS AND BAYESIAN DATA FUSION

4.1 Abstract

This paper explores the use of a deep convolutional neural network (DCNN) on mammographic images from the Digital Database for Screening Mammography (DDSM). Each case within the DDSM is either classified as benign or malignant (cancer) and has several views of the mass in question. The softmax output from the network is further processed in this work by first averaging the values for each image based on an individual. The overall value produced is weighted by an optimal cost factor based on minimum cost Bayesian hypothesis testing. Non-fused testing accuracy was 70% which was bolstered using fusion to 79%, yielding a 9% increase in testing classification accuracy. The results presented herein are preliminary and will be updated upon further investigation.

4.2 Introduction

Breast cancer is the most common form of cancer. It afflicts approximately 15% people worldwide and 12% here in the United States. The current mortality rate for breast cancer is 17%. Compare the overall mortality rate with 61% in the final stages and 99% in the initial stages. The stage at which the breast cancer is first detected directly correlates to
patient survivablility as well as quality of life. [22] The first three stages of breast cancer provide insight to the cancer’s progression and spread within the breast tissue. If caught at these stages, preferably stage one, the cancer is typically able to surgically removed. Unfortunately, not all cases are caught prior to their progression to stage four, the final stage, where the cancer is said to have metastasized, or spread to the surrounding organs and tissues. If caught at stage four, the patient will likely undergo unpleasant chemotherapy to treat the illness throughout the body [69].

Over the past several decades, digital image processing techniques have become instrumental in detecting a variety of cancers, such as prostate [39], colon [73], and brain [33]. Probably the most widely used case is mass detection on digital x-ray images. Mass detection can prove difficult due to the variety in mass size, shape, proximity to organs. A great deal of research has been done suggesting that benign masses tend to be more round, larger, with well defined margins. In contrast, malignant masses tend to be irregularly shaped, various sizes, with ill-defined margins [66, 10]. Image processing techniques seek to exploit the morphological differences to aid in patient diagnosis.

Early techniques in mammography used traditional image enhancement techniques coupled with feature extraction. The extracted features were then subjected to some form of supervised learning techniques such as a support vector machine [45]. Other approaches used included clustering techniques such as the fuzzy c-means to identify potential masses on images [3]. Some techniques took mass detection to the next step by classifying mass that were found to be either malignant or benign using a spiculation detection method [8].
Convolutional neural networks have seen a dramatic increase in application, specifically in image processing problems due to their ability to iteratively discover and learn features from images without user guidance. Specifically in mammography, the ability of CNNs to learn distinguishing features of malignant and benign masses is desirable due to the noisy and often highly variant breast tissue. The differences in breast tissue can cause traditional hand crafted feature-based discrimination methods to ill perform.

The accuracy of neural networks used for cancer detection in mammographs hovers around 85% with many attempts to wring out slightly better results. Some have tried combining traditional image processing techniques with neural networks [81, 60]. The work required the artificial expansion of the DDSM. The authors artificially expanding the DDSM from 1200 cases to 1.4 million test images by performing a multitude of cropping, scaling, inversion, and translation. With the larger amount of training data created, they were able to train a DCNN with a higher degree of success [70].

Several of the most successful approaches utilize transfer learning to overcome the lack of available training data. Huynh et al. utilize an a pre-trained CNN and compare its results with that of an SVM [41]. They reported an ROC of 81% for the SVM and 86% for the CNN using only 216 mammogram cases. The most significant result from this study is that the deep learner was able to outperform the hand crafted features even with such a low amount of training data.

Unsupervised approaches have also moved toward deep learning techniques. Ertosun et al, combine an unsupervised engine which utilized a R-CNN for localization of masses and a CNN for discrimination. [27] The two networks where trained using different training
The discriminator CNN is used to determine if an image likely contains a mass, if so it is sent to the R-CNN to draw a bounding box around suspicious mass. This study is interesting because it tries to do both localization and discrimination tasks together. The authors achieved a successful mass localization ROC of 85%, and 0.9 false positives per image.

The current best performing CAD mammography classification scheme is described in Carneiro et al.’s 2015 work. They take two pre-trained CNNs and further train them on MLO and CC views independently [16]. They achieved a 0.90 ROC on the DDSM as well as 0.90 on the InBreast dataset. This work is particularly exciting because it does not co-register the images before input into the network.

Common CNN methodologies combine viewpoints by finding landmark structures within images and the relative positions are ancillary input [2]. The network can utilize the relative positions to render a 3D image which can serve as a final system output or is an intermediate step in a larger machine learning application [61]. On the other hand, when relative position is not a necessary output CNNs have been shown to perform an internal co-registration of multiview images [47].

The Digital Database for Screening Mammography (DDSM) is the largest repository for mammographic images available [35]. It contains 697 normal cases, 916 malignant cases, and 867 benign cases. Each case contains a crano-caudal (CC) and a mediolateral-oblique (MLO) view for each breast, resulting in four images per case. Within the malignant and benign case files contain an overlay which encircles the region of interest (ROI) for screening. The ROI contains the suspicious mass at differing levels of precision. In fact,
the entire database is quite variant considering many doctors contributed to its creation. In fact, the mammograms vary widely in size, contrast, and content. This represents a new form of noise introduced into the system. To combat this noise, a subset of the database was generated by creating sub-images for each image with an ROI. These sub-images contain some normal tissue and some tissue of interest. These sub-images become inputs for the network. In total there are 6,657 malignant sub-images and 5,531 benign sub-images used in this project.

Bayesian statistics has long been utilized to boost detector performance. For instance, some CNNs have been included a Bayesian cost scheme as part of the optimizer [6]. Another study used a CNN to classify breast tissue and location as inputs into a Bayesian classifier for mass detection [79].

The cost value from the Bayesian scheme helps to form an overall case decision for the individual. Manipulating the cost for each class allows for scheme tunability. If the cost of misclassifying one class far exceeds that of another class then incorrectly choosing the penalized class requires the CNN to indicate a high degree of association with the first class to overcome the penalty. Although basic, the introduction of these cost values allows the user to maximize expectations with results. Furthermore, the model becomes more robust as small variations in results from the CNN can be reliably offset with a slightly higher cost value.

In this case, the binary detector is defined in eq. 2.1, where \( y \) is the observation vector of arbitrary size. In this instance, \( y \) will be the average value from the softmax output. The class decision is made based on the probability of the observation \( y \), weighted by the
cost of choosing this class incorrectly. Since there are only two classes in this case, the ratio will be sufficient to express the costs $C_{10}$ and $C_{01}$. There is a temptation to include the *a priori* probabilities into the detector which likely would increase model performance. However, this seems disingenuous, given that while the estimated prior probabilities are known, in a real use case, these would be unknown and so the model will be built to reflect the reality. Moreover, in real life, the *a priori* probabilities are very small, likely inhibiting the system. The costs are included in the detector to allow for model tunability. So, the $P_D$ will be maximized at the expense of a higher $P_{FA}$.

In application, the cost of a false benign is much greater than the cost of the false malignant case. In other words, the cost of incorrectly telling an individual that they do not have cancer is harmful, even fatal, since the misclassified cancer would then be permitted to continue its progression untreated. On the other hand, indicating to an individual that a mass is potentially cancerous when in fact it is benign causes the person to undergo an unnecessary biopsy. Although both of these outcomes are undesirable, classifying malignant masses as benign obviously outweighs the alternative. Therefore in practice, the cost of incorrectly choosing a malignant mass should be significantly higher.

4.3 Methodology

Due to limited availability of data, this project relies heavily on the use of data augmentation techniques. The first augmentation step was to artificially expand the database by cropping each mass from the raw mammogram from various positions in the output image. For example, a single image mass could be positioned in the bottom left, top right,
etc. on the output image. Each image generated in considered a unique instance for input into the CNN. For testing and training purposes, each case is either fully incorporated into the training or testing dataset ensuring that the trained CNN will never have seen the mass under test to give a more accurate sense of generalizability.

The output image is originally a $2048 \times 2048$, padded as needed, representation of the mass and is subsequently down-sampled into a size of $256 \times 256$. The down-sampling process serves two important functions. First, the raw images have a large variance of noise present and by down-sampling, much of the noise present is removed. Second, the smaller images allow for significantly faster training time as well as increased network complexity.

The convolutional network had eight convolutional layers. All convolutional filters had a kernel size of $3 \times 3$. Additionally, every convolutional layer has an exponential linear unit (ELU) activation layer immediately after. Using ELU for activation has recently shown promise in decreasing training time. In principle, the ELU differs from normal ReLU activation by using an exponential rather than a linear function, thus ensuring gradient flow.[21] After the activation layer a batch normalization was included to allow for gradient flow regardless of network depth. Between every two convolutional layers, a max pooling layer reduced the size of the feature maps from the previous layer by $1/4$. At the last convolutional layer, the feature maps, now reduced to a size of $8 \times 8$ are inputs into a fully-connected layer of 2,048 neurons. After the first fully connected layer, there are two more fully-connected layers which are of size 2048 and 1024, respectively. All fully-connected layers were subjected to dropout at a rate of 0.5. The network was trained until it converged.
to a training loss of 0.01 as shown in 4.3. In most instances, the network took anywhere from one to five days to train.

![Figure 4.1 Plot of the loss reducing with respect to the number of epochs.](image)

The results for each image within the testing data were recorded. The results for multiple images with multiple views for an individual were averaged together. For example, the MLO and CC view softmax output for the network would be averaged to produce a single value for the detector. The average result was used to make a Bayesian detector with unknown priors. The cost ratio between the two classes was swept to find the best accuracy attainable on a case basis.
4.4 Data

The DDSM is an amazing resource for CAD mammography but requires a significant amount of pre-processing prior to actual implementation into the network. A raw mammogram is shown in 4.4. The raw images are also intractability large to be input into a CNN as is as well as containing some varying amount of text. Because of these image inconsistencies, the raw images must first be downsized to increase the ease of network input. The right side of the figure refers to the overlay file which provides input from a radiologist about the location of the potential mass, hereafter referred to as the region-of-interest (ROI). In some cases, there are multiple masses described within the overlay file, each with its own unique features. In addition to providing the circumscription of the mass in each mammogram, the radiologist also offers the BI-RADS rating for each mass contained in the overlay file.

BI-RADS stands for Breast Imaging Reporting and Data System. It was developed to allow for categorization of breast masses invariant between physicians. A histogram of the BI-RADS ratings for each case within the DDSM is shown in 4.4.

Here is a list of the BI-RADS ratings:

- 0-incomplete
- 1-negative
- 2-benign findings
- 3-probably benign
- 4-suspicious abnormality
- 5-highly suspicious of malignancy

The amount of available training data is often the limiting factor in CNN performance and mammography is no different. Although the DDSM is the largest publicly available
resource for CAD mammography, their still is insufficient data to train a larger CNN. Therefore this project utilizes a heavy amount of augmentation techniques to artificially expand the database. These techniques include but are not limited to random flipping, cropping, rotating, and brightness adjustments. The networks apparent database is about 1.4 million images. These augmentation steps occur just prior to input into the network and a randomly applied. The effect of these transformation steps are shown in 4.4.

The mammograms from the DDSM vary considerably with regard to image quality as well as overlay-to-mass proximity contained within the 4,024 overlay files. In some instances, there is visible text giving some patient information contained on the image itself including, patient age, physician descriptions regarding mass morphology, and BI-RADS ratings.

4.5 Results

Discrimination of the benign and cancer images was marginally successful. Figure 4.5 shows the entire testing data results. The blue markers represent benign cases while the red markers represent the cancer cases. From this figure, some of the cancer cases appear to be easy to identify while most of the testing data lay in a large unseparated region.

The raw softmax output refers to an image which is a subset of a mammogram which is a subset of a breast, which is a subset of an individual. In total, there were 1,283 images classified by the network. Some baseline statistics are shown below in Table 4.1. The unscaled accuracy is the untreated accuracy of the data with no cost manipulation. The
optimal cost relates to the best cost found to increase the accuracy which is itself represent
as best accuracy.

<table>
<thead>
<tr>
<th>Table 4.1 Relevant values for raw softmax data.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of instances</td>
</tr>
<tr>
<td>Unscaled Accuracy</td>
</tr>
<tr>
<td>Optimal Cost</td>
</tr>
<tr>
<td>Best Accuracy</td>
</tr>
<tr>
<td>( P_{FA} ) at a ( P_D ) of 0.9</td>
</tr>
<tr>
<td>AUC at a ( P_D ) of 0.9</td>
</tr>
</tbody>
</table>

The fused softmax data showed encouraging results. The newly generated statistics are shown below in Figure 4.2.

<table>
<thead>
<tr>
<th>Table 4.2 Relevant values for fused softmax data.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of instances</td>
</tr>
<tr>
<td>Unscaled Accuracy</td>
</tr>
<tr>
<td>Optimal Cost</td>
</tr>
<tr>
<td>Best Accuracy</td>
</tr>
<tr>
<td>( P_{FA} ) at a ( P_D ) of 0.9</td>
</tr>
<tr>
<td>AUC at a ( P_D ) of 0.9</td>
</tr>
</tbody>
</table>

From the above table, the change in overall performance can be measured. The number of cases was reduced by a factor of ten, which is perhaps agnostics in terms of accuracy however does potentially introduce some non-linearity to the output graph. Without performing any cost sweeping analysis, the unscaled accuracy increased 5%. Once treated, the best accuracy also increased by 9%. The detector was able to achieve a decent probability of detection while not having excessive amounts of false alarms. The trends in accuracy are shown below in Figure 4.5.
4.6 Discussion

The approach used in this paper has several important points to consider. Using current best practices on raw cropped mammograms containing masses, the CNN described is unable to mostly unable to discriminate between images with cancerous or benign masses. Considering individuals using multiple images can boost performance overall and with tuning could produce consequential results.

Considering the relative simplicity of the proposed fusion-detection scheme, the results speak volumes on the impact post-processing techniques can have on seemingly poor raw results. With absolutely no changes to the CNN, the post-processing added an additional 9% to testing accuracy. Furthermore, the investigation performed over the duration of this work gives new and important insight to future data acquisition and processing.

Most importantly, this work suggests that a softmax output on any one image can be misclassified due to noise and perturbations within the image, but multiple images of the same mass but in different spatial locations can have a synergistic denoising effect based on the increase in accuracy. The results also suggest that not all images are created equally. In direct opposition, mammograms with more than one image extracted, the average variance was 0.7 and there was no significant correlation between an increase in image extracted number and correct classification accuracy. These conflicting results suggest a more nuanced effect in image number. All things considered, an increase in performance observed in this work seems worth the extra computation and processing.

In the future, a 10-fold cross-validation will further investigate the robustness and generalizability of this scheme. With only one trial conducted it is possible that some of the
trends observed are caused by artifacts within the data, such as random selection of data that is easily distinguished in testing/training can create a large variance in an individual fold result. Although the simplicity of this method is elegant, there are more powerful and proven techniques which could be compared to gauge relative suitability. An interesting candidate is the application of Choquet integrals to intelligently weight cases based on learned features [65, 7].

This work also lends itself to improving image extraction and selection. One possibility is to include a new “garbage collection” class which would have no mass present. This class could allow the network to converge quicker on the distinction between malignant and cancer classes. A similar approach involves adding a new network which determines whether a mass is in the foreground or background of the image. Any background image would not contribute to overall classification decision for the case.
Figure 4.2 LEFT: Raw mammogram of left breast (CC view) from cancer case 1132 from the DDSM. RIGHT: The same image with the mass area circumscribed.
Figure 4.3 Histogram of BI-RAD ratings for DDSM cases

Figure 4.4 LEFT: Cropped image taken from the case in 4.4, CENTER AND RIGHT: Two examples of training images after the augmentation step in the network
Figure 4.5 Softmax output from CNN, each axis represents the networks determination of likeness associated with each class.
Figure 4.6 Accuracy as plotted as a function of the scaling cost. The unfused results are colored in blue were the fused results are colored in red. The x-axis is logarithmically scaled.
CHAPTER 5
CONCLUSIONS

5.1 Contributions

This thesis contributed two novel methods for the detection of breast cancer in digital mammograms. While these methods are not limited to the digital mammography field, they are specifically tuned to fit the dataset heuristics. Furthermore, newer CNN elements are implemented and compared to their traditional counterparts.

The first method handles the multiple input problem differently. Rather than train the model on a single image at a time, the CNN was trained with multiple views simultaneously, a multi-view input. In addition, new network elements such as SELU and ELU were employed and their effects on testing accuracy were explored. Lastly, the multi-view model uses several image processing techniques to produce novel input images which can be augmented during training to reduce over-fitting.

The second method adds a Bayesian decision-level fusion technique on top of a typical CNN model to unify multiple results and improve testing accuracy. This method was developed with screening applications in mind where physicians often take multiple images of one or more masses from different angles per breast. Considered alone, the classifier struggled to correctly identify mass malignancy but with the inclusion of multiple images for a patient, the system performance improved significantly.
5.2 For Further Research

The research herein is currently cutting-edge and so lends itself to a great deal of further experimentation. Future work for the Bayesian fusion model includes: (1) include decision-fusion into the network rather than after completion, (2) fuse outputs from several CNNs in parallel using different processing techniques, (3) include ancillary data into fusion model using an SVM.

In regards to the mutli-view model, future work will involve: (1) implementing a GAN to further increase apparent training data size, (2) apply a recurrent or region based CNN archetype, (3) construct a three dimensional CNN which classifies a voxelized representation of the breast tissue, (4) using successive autoencoders rather than hand-crafted image processing techniques to increase training data size. As the DL research community flourishes, new network architectures are presented, all of which could be applied to mammographic data.
REFERENCES


[81] I. Wichakam and P. Vateekul, “Combining deep convolutional networks and SVMs for mass detection on digital mammograms,” 2016 8th International Conference on Knowledge and Smart Technology (KST). feb 2016, pp. 239–244, IEEE.


