A Comparison of Deep Learning Algorithms for Medical Image Classification and Image Enhancement

Carina Pereira

A thesis

submitted in partial fulfillment of the

requirements for the degree of

Master of Science in Bioengineering

University of Washington

2018

Reading Committee:

Michalakis Averkiou, Chair

Adam M. Alessio

Program Authorized to Offer Degree:

Bioengineering

University of Washington

Abstract

A comparison of deep learning algorithms for medical image classification and image enhancement

Carina Pereira

In recent years, machine learning techniques based on neural networks have gained popularity. This is primarily because of improved computational capabilities and the availability of larger datasets. In our work, we investigate the application of machine learning techniques, specifically Convolutional Neural Networks (CNNs), for the purpose of medical image analysis. We consider three different tasks for our analysis. Two of these are classification tasks and the third task is an image enhancement task. In the first task, we classify thyroid nodules as malignant versus benign on B-mode and Shear Wave Elastography (SWE) images. We obtain accuracies ranging from 80% - 87% using our evaluated approaches. In the second task, we automatically classify breast Magnetic Resonance Imaging (MRI) images into lesions present and lesion absent classes. For this task, we obtain accuracies ranging from 56% - 69%. In the third project, we train and evaluate a deep learning algorithm for up sampling low resolution ultrasound images and present promising results for obtaining high resolution images from lower quality acquisitions. In general, this work demonstrates that models reliant on deep learning with 10^4 to 10^8 unknown parameters can be trained and effectively applied with modest data set sizes on the order of 500 to 10,000 images.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would first like to express my gratitude to my thesis advisor Adam Alessio for his support and kindness. Adam has been an incredible mentor; I am grateful for the interest he has shown in my career. Whether I've found myself in trouble or I want to discuss research questions, his office door has always been open. He has helped me make the most from my time at the University of Washington. Thank you for all your time and effort. I would also like to thank Mike Averkiou for chairing my committee. Thank you for your insightful questions and providing helpful feedback.

I would also like to thank my advisors Dr. Manjiri Dighe, Savannah Partridge and Thanasis Loupas for all their help and support these past two years. Thank you for your insightful comments and encouragement. I would like to thank Jeff Thiel for collecting and organizing the data for our thyroid nodule classification problem. I would also like to thank Yifan Wu for taking the time to answer all my questions and organizing data for the breast lesion classification project. Special thanks to Ginger Lash for finding artifacts in the DCE-MRI images. I would also like to thank Unmin Bae and Philips Healthcare for allowing me to continue the deep learning based upsampling project after the completion of my internship.

For the last two years I've had the pleasure of working with different people from the Imaging Research Lab (IRL). I would like to thank Achille Mileto, David Zamora and Kalpana Kanal for allowing me to work on the CT detectability project. Thank you for all your advice and support. I would also like to thank Paul Kinahan and Ruoqiao Zhang for teaching me about CT image reconstruction. I would like to thank Tzu-Cheng (Efren) Lee, Sandra Johnston, Chengeng Geng, Nathan Bell, Chris Sanchez, Adrienne Lehnert, Vivi Wu, Peter Muzi, Robert Harrison and Michael

Bindschadler for your wonderful company in the lab. It has made the last two years so much more fun and exciting.

Our research work would not have been possible without the use of the CLEAR server. Thank you, Nathan Cross, for allowing me to use it. Special thanks to Matt Bruce for his insights on the Supersonic ultrasound system. I would also like to thank all the program advisors at the Bioengineering department: Chetana Acharya, Peggy Sharp and Kalei Combs for their support and encouragement.

Finally, I would like to thank all my friends and family for their patience and never-ending support. I would not have been able to do any of this without your love and support.

Chapter 1. BACKGROUND

Machine learning algorithms can learn tasks by generalizing patterns from data [1]. This often makes performing complex tasks cost effective and feasible when manual programming is not. Examples of such complex tasks include speech recognition, fraud detection and stock market prediction [2]. Machine learning techniques can be easily applied to certain fields such as radiology, which mostly relies on extracting useful information from medical images [3]. One pre-requisite for applying machine learning techniques is the availability of a large data set. Currently, 30% of the world's stored data is generated in the healthcare industry [4]. Recent advances in machine learning coupled with the availability of medical images has resulted in the development of different machine learning algorithms for medical image analysis. These include the detection and classification of breast lesions from mammograms [5], segmenting anatomical structures from Magnetic Resonance Imaging (MRI) [6] and metal artifact reduction in Computed Tomography (CT) images [7].

1.1 INTRODUCTION TO NEURAL NETWORKS

Neural networks are a type of machine learning algorithm that have gained popularity in recent years [8]. Although the idea of neural networks has been in place for decades, they have only recently gained popularity due to the availability of large data sets and improved computing capabilities.

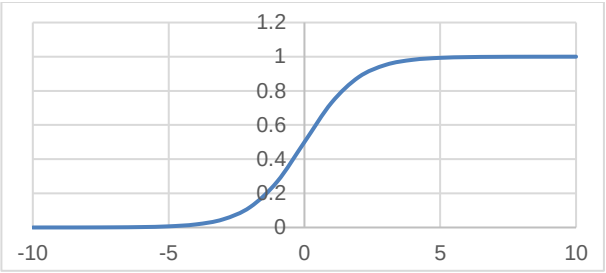
1.1.1 *History of neural networks*

Neural networks as we know of them today have come a long way from their origins in the area of human brain modeling. McCulloch and Pitts (1943) initially developed a model of how neurons

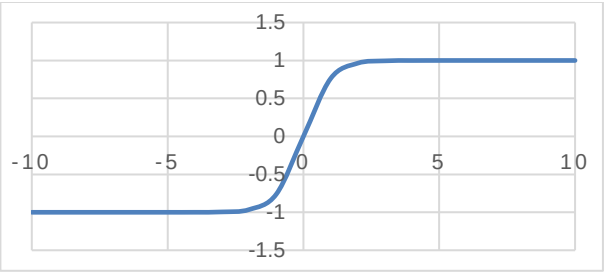
function in the brain. This model is considered to be the ancestor of the neural network [9]. Donald Hebb (1949) later published a book called 'The Organization of Behavior' where he introduced what was later known as Hebbian learning. It was one of the first learning rules for neural networks [10][11]. Frank Rosenblatt (1958) later introduced the first perceptron, modelled on the McCulloch-Pitts neuron. This perceptron had the ability to converge to the correct solution and learn weights for the problem at hand. The perceptron quickly gained popularity, but this came to an end in 1969 when Marvin Minsky and Seymour Papert argued in their book 'Perceptron' that the single perceptron model could not be translated into multi-layered neural networks. This ended a lot of research into neural networks and the period soon after is often known as the "AI winter" [9]–[11]. Much later, Rumelhart, Hinton and Williams (1986) published a paper titled, 'Learning representations by back-propagating errors' [12]. In this paper they showed that neural networks with many layers can be trained by a simple procedure. This led to a lot of excitement in the field and eventually gave us neural networks [9], [11]. There is still some controversy however as to who should be credited for some of these earlier developments [13]. It took another 10-15 years before data sets became large enough and computing power improved for neural networks to become as prevalent as they are today [10].

1.1.2 *What are neural networks?*

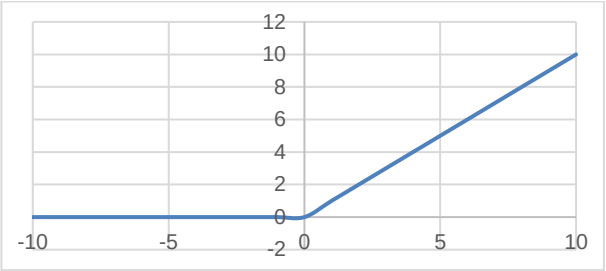
A neural network consists of neurons that have learnable weights and biases. Each neuron in the network performs a dot product operation between the input to the neuron and its weights plus a bias. The result of this operation is then fed to a non-linear/activation function. These activation functions are what allow the neural networks to learn non-linear decision boundaries [14][15]. Some commonly used activation functions like the sigmoid function, the tanh function and the Rectified Linear Unit (ReLu) function are illustrated in Figure 1.1.



(a) Sigmoid function



(b) Tanh function



(c) ReLu function

Figure 1.1. Commonly used activation functions. (a) Sigmoid function (b) Tanh function (c) ReLu function

Activation functions are important because without them the neural network would only be able to learn a linear decision function [15]. Currently, ReLu functions are popular because of their ability to handle the vanishing gradient problem [15]. The standard equation of a neuron is,

$$y_i = f\left(X_i \bullet W_i + b_i\right) \tag{1.1}$$

where X_i are the inputs to the i^{th} neuron, b_i is its bias, W_i are its weights, f is a non-linear function and y_i is the output of the i^{th} neuron. The neural network has a series of layers where each layer consists of neurons that are connected to the output of the previous layer as depicted in Figure 1.2., where x_{11}, x_{21} and x_{31} are the inputs to the neuron, W_1 are its weights, b_1 is its bias and y_1 is the output of this neuron which also serves as input to the next layer. For a classification problem, the last layer of a neural network gives us a score indicating the likelihood that the input data belongs

to a certain class. The architecture of a neural network illustrated in is for a two-class classification problem.

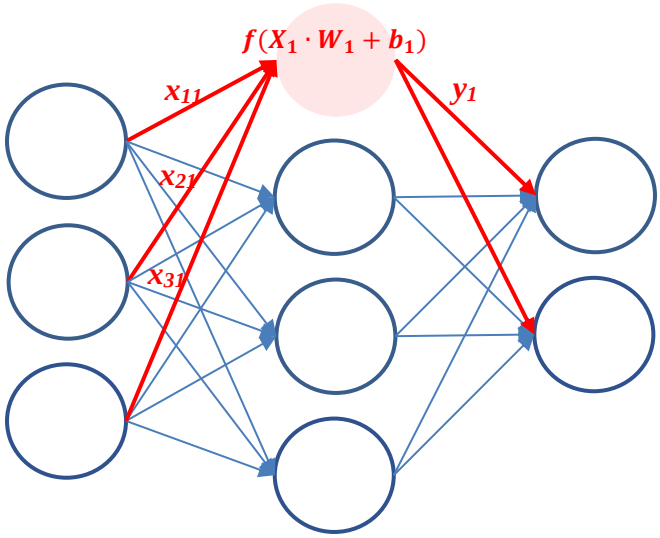


Figure 1.2. A two-layer neural network.

1.2 CONVOLUTIONAL NEURAL NETWORKS

A type of neural network called Convolutional Neural Networks (CNNs) are generally used for computer vision tasks. These neural networks take advantage of the fact that the input is an image and there exists some relationship between pixels in that image [2][14]. CNNs have different types of layers including convolutional layers, pooling layers and fully connected layers among others.

The key layer of a CNN is the convolutional layer. The convolutional layer has neurons arranged in a three-dimensional manner (width, height and depth), that are called convolutional kernels/filters. Depth in this context means the number of filters and not the number of layers in the neural network. These convolutional filters are convolved with the image volume, these are slid across its width and height dimension. The output of this operation gives us an activation map that tells us the response of the filter at every spatial position. Every convolutional layer has a fixed number of filters, the activation map generated by every filter is stacked together along the depth

direction. The output of the convolutional layer is typically followed by an activation layer or a non-linear function (eg. ReLu, sigmoid or tanh function). Intuitively, the network learns filters that are activated by certain features like edges, colors or shapes [14].

Pooling layers are commonly inserted between convolutional layers, these reduce the spatial size of the input data without affecting its depth. An example of a pooling layer is a 2x2 max-pool layer, these work by retaining only the maximum value of every 2x2 section of its input volume, thus reducing the spatial size of the input volume by 50%. This is done to reduce the number of parameters that the network has to learn and acts as a control for overfitting [14].

A fully connected layer for a CNN is one where every output of the previous layer is input to the fully connected layer. For a classification problem, the last layer of a CNN can be a fully connected layer that outputs a score for each one of its classes. It has learnable weights and biases, it has the same functionality as the layers found in a neural network previously described in Section 1.1.2.

Other layers like the drop out layer, the batch normalization layer and the transposed convolutional layer can also be used in a CNN. A typical CNN architecture is depicted in Figure 1.3, it consists of two convolutional layers, one max-pooling layer and a fully connected layer. Like a regular neural network, the final layer in a CNN contains N scores, where N is the desired number of different classes. The scores represent the likelihood that the input belongs to each class. A final step imposes a threshold (depending on desires sensitivity/specificity) on these class scores to determine the final class.

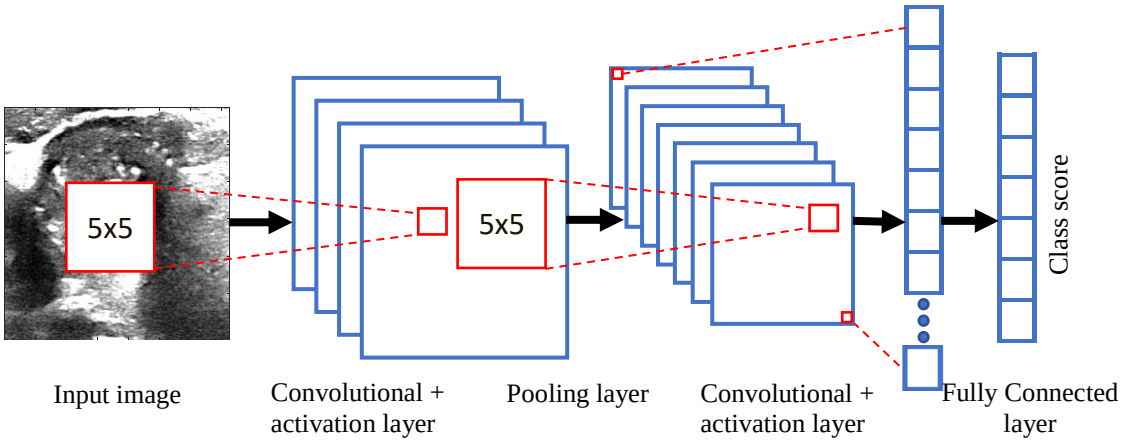


Figure 1.3. A CNN with two convolutional layers, a max-pool layer and two fully connected layers

1.3 TRAINING A CNN

CNNs vary in terms of complexity and the number of trainable parameters. Choosing the right network architecture differs depending on the problem at hand. Besides selecting the right architecture, we must also make sure that the CNN learns weights that would give the best performance on any new data set.

1.3.1 *Loss functions*

The training objective of any neural network is to minimize the discrepancy between the predicted output and the true output. The error/loss function, $E(W,b)$ is generally used to measure this discrepancy. Here, W and b are the learnable parameters of the neural network. The selection of the loss function is an important algorithm decision and should depend on the problem and data at hand. Commonly used loss functions are the mean squared error and the cross-entropy loss [14], [16], [17]. Cross-entropy loss for a two-class classification problem is given as,

$$E(W,b) = -(y\log(p) + (1-y)\log(1-p)) \tag{1.2}$$

where y is the true class of the input image and p is the predicted class score from the CNN. The mean square error loss described in Eq 1.3 is another loss function. This loss function is commonly used in CNNs trained to generate high resolution images from low-resolution images. The mean square error loss function is given as,

$$E(W,b) = \frac{1}{n}\sum_{i=0}^{n}\left(F\left(Y_i;W,b\right) - X_i\right)^2 \tag{1.3}$$

where Y_i is the low-resolution image, X_i is the high-resolution image, $F(Y_i;W,b)$ is the image generated from the neural network and n is the number of samples.

1.3.2 *Optimization algorithms*

It is not possible to find an analytic solution for the weights of a neural network because the loss function is a non-convex function i.e. the loss function has multiple local minima in the parameter space [16]. Instead we use an optimization algorithm that finds weights which minimize the loss function. The optimization algorithm is an iterative process of the form,

$$W^{k+1} = W^k + \Delta W^k \tag{1.4}$$

where k is the iterative step. Different optimization algorithms can be used to find the update value ΔW^k, the most basic of these being the gradient descent algorithm. The gradient descent algorithm is of the form.

$$W^{k+1} = W^k - \alpha\nabla E\left(W^k\right) \tag{1.5}$$

where α is the learning rate and $-\nabla E(W^k)$ is the direction of the greatest rate of decrease of the error function. Eq. 1.5 shows that the weights are updated in an iterative manner. It eventually reaches the local minima as illustrated in Figure 1.4. Note that the learning rate, α is an important

hyperparameter that controls the rate of convergence. A learning rate that is too small means that we have slow convergence while a learning rate that is too large implies that the algorithm never converges.

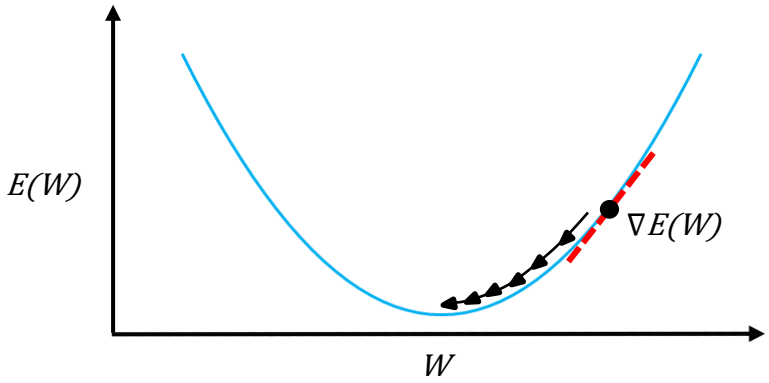


Figure 1.4. Illustration of the gradient descent algorithm for a loss function in a one-dimensional parameter space. Note, that the weights are updated in a manner such that it eventually reaches a minima.

The optimization algorithm is another important choice that can affect the performance of the network. Currently, the Adam optimization algorithm is commonly used as the default. It uses adaptive estimates of lower order momentums to update the weights [18].

1.3.3 *Training and testing data*

There tends to be three different data sets in machine learning: the training data set, the validation data set and the testing data set. The training data set is used for training purposes only; the weights of the neural network are optimized on this data set. The second data set is the validation data set, which is used to evaluate and select hyperparameters such as the optimization algorithm or the type of loss function used. The hyperparameters that give us the best performance on the validation set are eventually selected for our network. Once the network has been trained using the training data set and the hyperparameters have been selected using the validation data set, we evaluate the

performance of the network on the test data set. The test data set is an independent data set that the network has not seen during the training phase [16]. This is important because it provides us with an unbiased estimate of the model performance [19]. The performance of the network on each of the data sets is important: the performance of the network on the training data tells us if the network has enough learnable parameters, the performance of the network on the validation data set tells us if the network is overfitting and the performance of the network on the testing set tells us about its generalizability i.e. its performance on an independent data set.

1.4 HISTORICAL CONTEXT OF CNNS FOR MEDICAL IMAGING TASKS

Neural networks have been applied to medical imaging tasks for decades [20]. Lo et al. used a CNN to identify lung nodules using chest radiographs in 1993 [21]. Chan et al. used a CNN to identify microcalcifications on breast mammograms in 1995 [22]. Sahiner et al. used a CNN to characterize malignancy on breast nodules using breast mammograms in 1996 [23]. All three of these CNNs had only two convolutional layers. The architecture of these CNNs were limited due to the computationally expensive process of training CNNs in the pre-GPU era [20]. GPUs can perform matrix and vector multiplications a lot faster than CPUs. Since this is a basic task in training a neural network, GPUs have made training CNN's more efficient, enabling the use of more complex architectures [24]. Training CNNs is now approximately 40 times faster compared to using a CPU [20][24].

Deep neural networks are currently popular. It is a type of neural network consisting of many layers (>5) that can extract high level information from input data [20]. Deep CNNs have become powerful tools that automatically learn high and mid-level abstractions from images automatically [20]. CNNs have been developed for different medical imaging tasks like

segmentation [6], classification [5] and artifact reductions [7]. Training a deep CNN from scratch, however, can be challenging. These challenges are described in the following section.

1.5 CHALLENGES WITH TRAINING CNNS FOR MEDICAL IMAGING TASKS

1.5.1 *Limited medical imaging data*

A pre-requisite to training a CNN for any medical image analysis task is the availability of a large high-quality imaging data set. The majority of machine learning applications in medical imaging have relied on supervised learning approaches that require labeled reference data, where the images are labeled by an expert. This is often an expensive process that might require input from radiologists or pathology [25]. Due to the limited number of data points in most medical imaging tasks and the large number of weights that need to be learned, CNNs can easily overfit i.e. it learns weights that yield good performance on the training data set and poor performance on the validation/testing data set. One way to check for overfitting is to plot the loss of the training data set and the validation data set during the training process as illustrated in Figure 1.5. Note that the network has a much lower loss on the training data compared to the validation data. One way to reduce the effect of overfitting is to use an early stopping criterion i.e. if the loss on the validation data has not reduced after a fixed number of iterations then end the training process. Other methods exist to limit the effect of overfitting besides acquiring more data, including adding a drop out layer, regularizing the network and augmenting the images during the training process.

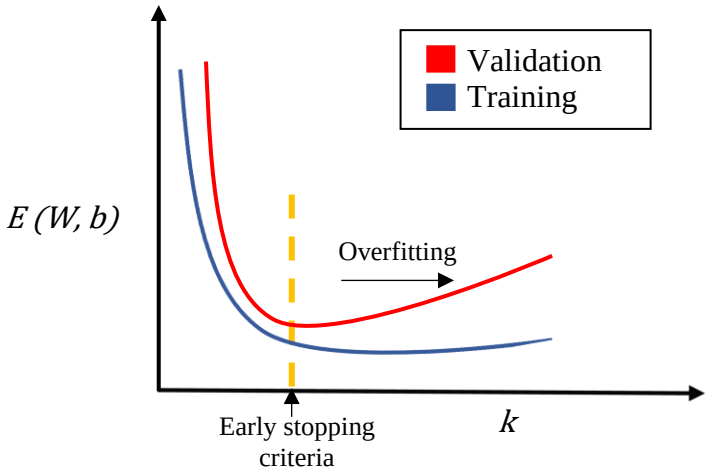


Figure 1.5. Illustration of a learning curve for an overfit CNN. After a certain number of iterations of the optimization algorithm, the performance of the CNN on the training data set improves while the performance of the CNN on the validation set decreases.

Adding a drop out layer to a CNN implies that during the training phase, the weights of a random fraction of neurons are set to zero [26]. We obtain the neural network in Figure 1.6 after adding a drop out layer to the neural network previously depicted in Figure 1.2. As illustrated about half the weights are randomly set to zero during every iteration of the optimization process. This allows the network to learn weights independently from each other and helps reduce overfitting. Note that during the evaluation phase, none of the weights are set to zero.

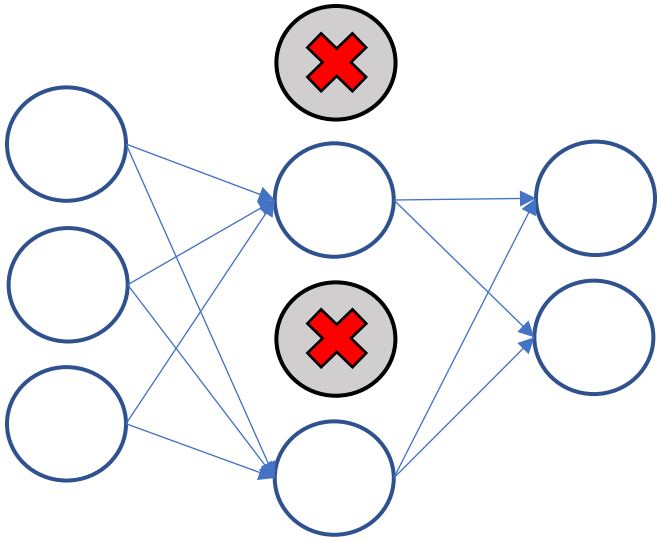


Figure 1.6. A regular neural network with half the weights in the hidden layer set to zero

Regularization is another way to prevent overfitting. It adds a penalty term to the loss function and forces the network to learn smaller weights. The penalty term is another important hyperparameter which can affect the performance of the CNN.

Another way to reduce the effect of overfitting is through data augmentation. The size of the training data set is artificially increased by randomly cropping, flipping and scaling the images [27] or through more advanced synthetic data generation such as generative adversarial networks [28], [29].

1.5.2 *Unbalanced data set*

Another potential challenge to training a CNN for medical image analysis is the unbalanced nature of the medical data sets. Often machine learning algorithms assume that there exists equal number of data points per class. In the medical imaging field however, the number of diseased i.e. malignant cases are often much smaller compared to the non-diseased i.e. benign cases. This is problematic because the CNN would learn weights for a trivial classifier where it classifies all the

images as benign. One way to limit the effect of class imbalance is to associate a cost with misclassifying every image during the training phase. The cost of misclassifying a malignant case is higher than the cost of misclassifying a benign case. This forces the network to learn more from the malignant cases during the training process even if the number of malignant cases is smaller compared to the number of benign cases.

1.6 THESIS STRUCTURE

The purpose of this thesis is to investigate the application of machine learning techniques, primarily CNNs for medical imaging tasks. In our study, we have considered three different medical imaging tasks. Two of these tasks are image classification tasks and the third is an image quality improvement task. In the first task, we characterize malignancy in thyroid nodules using their corresponding Shear Wave Elastography (SWE) and B-mode ultrasound image. In the second task, we develop a breast lesion detection algorithm using their Maximum Intensity Projection (MIP) Magnetic Resonance Imaging (MRI) images. In the third task, we train and evaluate a Generative Adversarial Network (GAN) to generate a high-resolution ultrasound image from their corresponding low-resolution image.

In chapter 2, we compare the performance of different machine learning algorithms to characterize thyroid nodules. We utilize B-mode ultrasound images and their corresponding SWE images of thyroid nodules for our analysis. We compared a spectrum of machine learning approaches: feature extraction plus supervised learning, two fully trained and a finely tuned CNNs.

In chapter 3, we characterize breast lesions using their Maximum Intensity Projection (MIP) Magnetic Resonance Imaging (MRI) images. We train and evaluate different neural networks to detect the presence of a lesion using its MIP MRI image.

In chapter 4, we trained and evaluate the Super Resolution Generative Adversarial Network (SRGAN) to generate high resolution ultrasound images from their corresponding low-resolution ultrasound image. We evaluated the performance of the SRGAN and compared it to the standard bicubic interpolation.

Chapter 2. MACHINE LEARNED ALGORITHMS FOR AUTOMATIC ANALYSIS OF THYROID NODULES FROM SHEAR WAVE ELASTOGRAPHY IMAGES

Thyroid nodules are extremely prevalent and occur in about 50% of the population based on autopsy data [30]. However, only 3% to 7% of these are malignant [31]. Currently, the most definitive way to identify malignancy is through Fine-Needle Aspiration (FNA) or through surgery [32]. Since it is not practical to biopsy every single thyroid nodule to confirm malignancy, there is a need to noninvasively identify which nodules warrant the need for a biopsy. Ultrasound imaging is an important diagnostic tool used to assess malignancy in thyroid nodules [32], [33]. Several studies have been able to identify and evaluate sonographic features that relate to the likelihood of malignancy in thyroid nodules. Multiple classification systems and guidelines have been developed to solve this problem including the ones developed by Society of Radiologists in Ultrasound in 2005 and American Thyroid Association in 2015 [34]. In 2017, American College of Radiology (ACR) developed an ultrasound based risk stratification system called the ACR TIRADS (Thyroid Imaging, Reporting and Data System) [32]. Although commonly used, these sonographic features are not perfect predictors of malignancy and risk stratification remains challenging.

Shear Wave Elastography (SWE) has been used to predict malignancy in thyroid nodules by measuring tissue stiffness. A focused ultrasound beam is used to generate a vibrating source in the tissue [35]. These vibrating sources cause shear waves to propagate through the tissue resulting in tissue displacement that is then used to estimate the shear speed of the tissue. The stiffness of the tissue (in kPA) is obtained quantitatively from its shear wave speed [36], [37]. Studies have shown that tissue stiffness as measured by SWE can be predictive of malignancy since malignant

nodules are thought to be slightly stiffer than benign nodules [38], [39]. A recent meta-analysis has shown that clinicians are able to identify malignancy in thyroid nodules using elastography techniques with a diagnostic accuracy of 81.7% [40].

In this work, we develop different machine learning approaches for the detection of malignancy in thyroid nodules using its B-mode ultrasound image and SWE image with the goal of finding a fast, automatic strategy for accurately characterizing nodules. Preliminary results of this effort were presented previously [41]. In our work, we compared the performance of an automatic feature extraction plus supervised learning algorithm, two fully trained CNNs and a finely-tuned ResNet [42].

2.1 Literature Review

To facilitate an accurate and quick diagnosis of thyroid nodules, various Computer-Aided Diagnosis (CAD) systems have been developed [43]–[47]. Some of these CAD systems characterize the nodules by automatically extracting features from B-mode ultrasound images of the thyroid nodules [43]–[46]. Recent CAD systems utilize pre-trained CNNs to characterize thyroid nodules using its B-mode ultrasound image [47], [48]. Ma et al. used two fused pre-trained networks to characterize malignancy in thyroid nodules using its B-mode image [47]. The authors trained the network on about 15000 images of manually segmented thyroid nodules. They fused the output layers of the two pre-trained networks and made a prediction based on this fused output. Note that each network had its own unique architecture. The authors achieved an accuracy of about 83%. Chi et al. used a pre-trained network to extract features from B-mode images of the thyroid nodule. A random forest classifier was then used to predict malignancy using the features extracted from the pre-trained network. The network was trained on patches extracted from 592 images of thyroid nodules. These were delineated by radiologists. The authors achieved an accuracy of 96%

and an AUC of about 99%. Note that the authors did not define malignancy based on pathology but used the TIRAD scores instead [48]. Liu et al. used a hybrid approach whereby they combined features extracted from a pre-trained network and handcrafted features to predict malignancy in thyroid nodules [49]. They performed their analysis on a data set of 1037 delineated images of thyroid nodules. This method achieved an accuracy of about 93% and an AUC value of 97%. Most of the recent CNN based classification systems characterize malignancy based on the B-mode image of thyroid nodules. There has been limited work on the development of CAD systems that utilize SWE images for the analysis of thyroid nodules.

2.2 Dataset Description

In this study, SWE and corresponding B-mode images of thyroid nodules were obtained for 165 patients. The patients that participated in the study were scheduled to receive a thyroidectomy or FNA based on Society of Radiologist in Ultrasound (SRU) guidelines [35]. These images were retrospectively analyzed to evaluate the predictive ability of different machine learning algorithms to classify each image as malignant or benign. The pathology report served as the reference test to confirm malignancy.

For every patient included in the study, SWE and B-mode images were obtained prior to FNA or thyroidectomy. The SWE images were obtained on a SuperSonic Imagine's Aixplorer with a linear array transducer having a bandwidth of 4 to 15 MHz. The SWE images of thyroid nodules were acquired by seven sonographers trained in elastography for 5 years but still new to SWE imaging. The SWE image was acquired by placing the transducer lightly on the patients neck after the ultrasound gel was applied. A B-mode image of the thyroid nodule was displayed. The sonographer then drew a Region of Interest (ROI) around the thyroid nodule using the B-mode

image as reference following which a SWE image was overlaid on top of the B-mode image and displayed in a dual monitor setting. An example is provided in Figure 2.1.

The images in the data set were obtained - along the axial and transverse orientations, at different depths from the skin surface, and with different physical imaging fields of view (0.95 cm to 2.96 cm) at different pixel sizes. A total of 964 images were obtained, 752 of these belonged to benign nodules and 212 of these belonged to malignant nodules. These images were then subdivided into a training (82%) and a testing (18%) set. They were subdivided in such a way that the prevalence of malignancy was the same in both sets and that the images in the training set belonged to a different group of patients compared to the images in the testing set. This was done to ensure that the images in the training and testing data sets were independent.

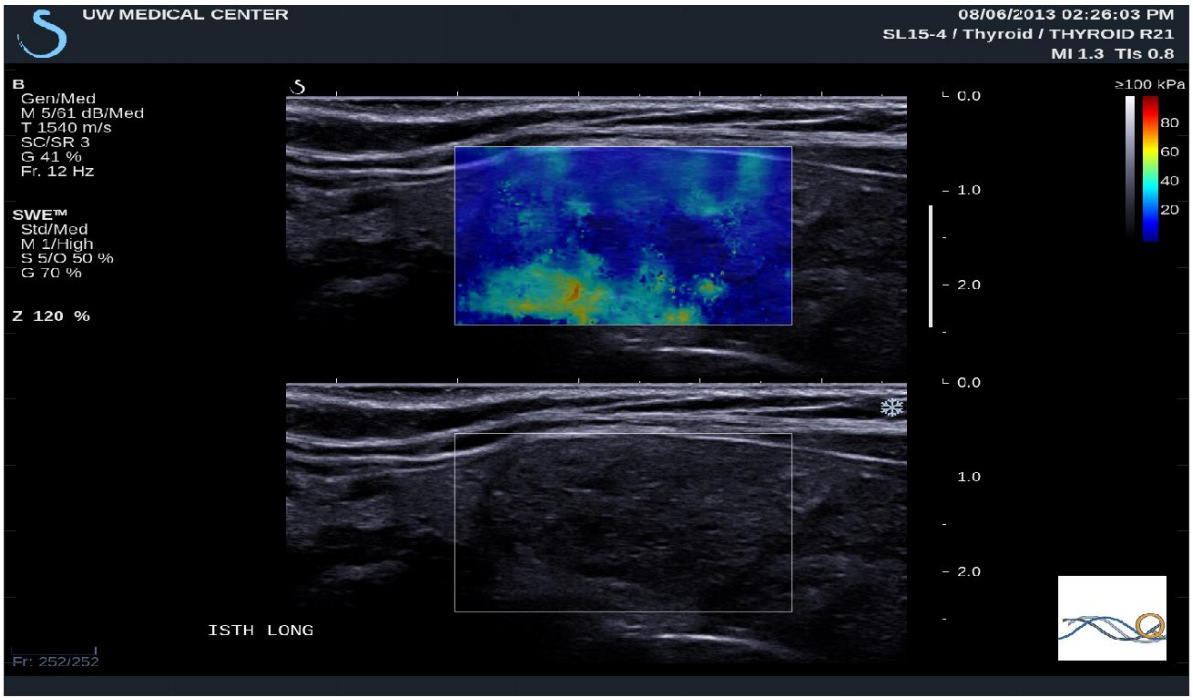


Figure 2.1. SWE image of the thyroid nodule overlaid on its corresponding B-mode ultrasound image. The position of the ROI within which the SWE image has been displayed was decided by a trained sonographer

2.3 IMAGE PREPARATION

The SWE and B-mode information of the thyroid nodules was stored in the DICOM image. An example of the DICOM image is provided in Figure 2.1. The upper half of the DICOM image contains the SWE image overlaid on its B-mode image while the bottom half of the DICOM image contains the B-mode image only. First, we segmented the blended image and corresponding B-mode image using the ROI defined by the sonographer. We assumed that alpha blending was used to blend the images. The equation for alpha blending is given in Eq. 2.1

$$I_{Blend} = (1-\alpha) I_{B\mathrm{mode}} + \alpha I_{SWE} \tag{2.1}$$

Where I_{Blend} is the overlaid image, $I_{B\mathrm{mode}}$ is the B-mode image, I_{SWE} is the SWE image and α is the blending factor. We obtained the SWE image from the blended image by assuming α to be 0.5. α was selected based on empirical analysis of all our images. For training and evaluating the pre-trained network, we used this pre-processed SWE image and corresponding B-mode image. For the other machine learning based classification algorithms, we followed an additional step.

After extracting the SWE image from the blended image, we mapped each of the RGB values in the image to its stress value in kPA using the colorbar present on the DICOM images. Note that all the DICOM images had the same colorbar and hence all the pre-processed images were mapped to the same range of stress values. We used this pre-processed SWE image and corresponding B-mode image for training and evaluating the simple feature extraction plus machine learning

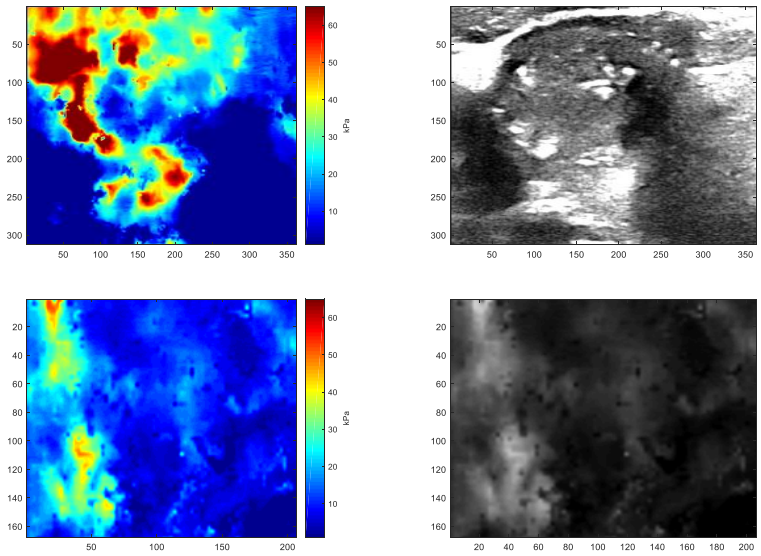


Figure 2.2. Example SWE and B-mode ultrasound image for (a) Benign nodules (top row) and (b) Malignant nodules (bottom row)

algorithm, the one class-auto encoder and the Siamese neural network. Example images used for our analysis are provided in Figure 2.2.

2.4 METHODOLOGY

In our work, we evaluated different machine learning algorithms for characterizing malignancy in thyroid nodules. We have compared the performance of an automatic feature extraction algorithm followed by different supervised learning algorithm, two fully trained deep learning algorithms and finely-tuning the ResNet50 [42].

2.4.1 *Simple feature extraction plus machine learning classification algorithm*

Conventional image classification algorithms work by extracting features from input images and subsequently feeding these features into a machine learning algorithm. Machine learning algorithms are able to generalize patterns from images automatically [1]. Feature selection is an important process that usually requires some domain knowledge. In our study simple features such as the mean, the standard deviation, the range of stress values, stress values greater than a threshold of 80 kPa and highest stress value were extracted from the SWE images. The circular Hough

transform was also used to extract features describing the circularity of the thyroid nodules. This feature is important because shape is a clinical factor used to evaluate malignancy of a thyroid nodule. These metrics were then fed into machine learning algorithms including Naïve Bayes algorithm, support vector machines (SVM's), logistic regression classifier and the decision tree algorithm. These algorithms were implemented on MATLAB [50]. To prevent the machine learning algorithms from learning the weights of a trivial classifier which classifies all the images as the majority class (benign nodules), we oversampled the minority class (malignant nodules) during the training phase. We oversample by repeating the instances of the malignant nodules such that its prevalence is about 50% of the training instances.

2.4.2 *Fully trained neural network*

One-class auto encoder

An auto encoder is a type of unsupervised neural network that learns encodings of the data. It is often used as a data compression algorithm. In our case, we used the auto encoder to learn encodings from the images of the benign thyroid nodules only. This approach is based on the logic that the auto encoder learned the encodings of the benign images only and therefore should be able to reconstruct the benign images with a low mean squared error (MSE), computed between the reconstructed image and the original image. When we use the same auto encoder to reconstruct the malignant images, we expect these images to have a higher MSE. Consequently, the MSE can be used to classify the image as benign or malignant. This approach has been previously used to identify malignancy in breast images [51].

The auto encoder consists of an encoder and a decoder portion. Its architecture is depicted in Figure 2.3. The architecture was based on the autoencoder implemented using Keras in [52]. In total, the network has 4,385 trainable parameters. The number of trainable parameters in our one

class auto-encoder is much smaller compared to state of the art convolution neural networks such as AlexNet[53] or VGG16 [54]. The lower number of trainable parameters in the one-class neural network should help provide reasonable performance with limited training data and help reduce overfitting.

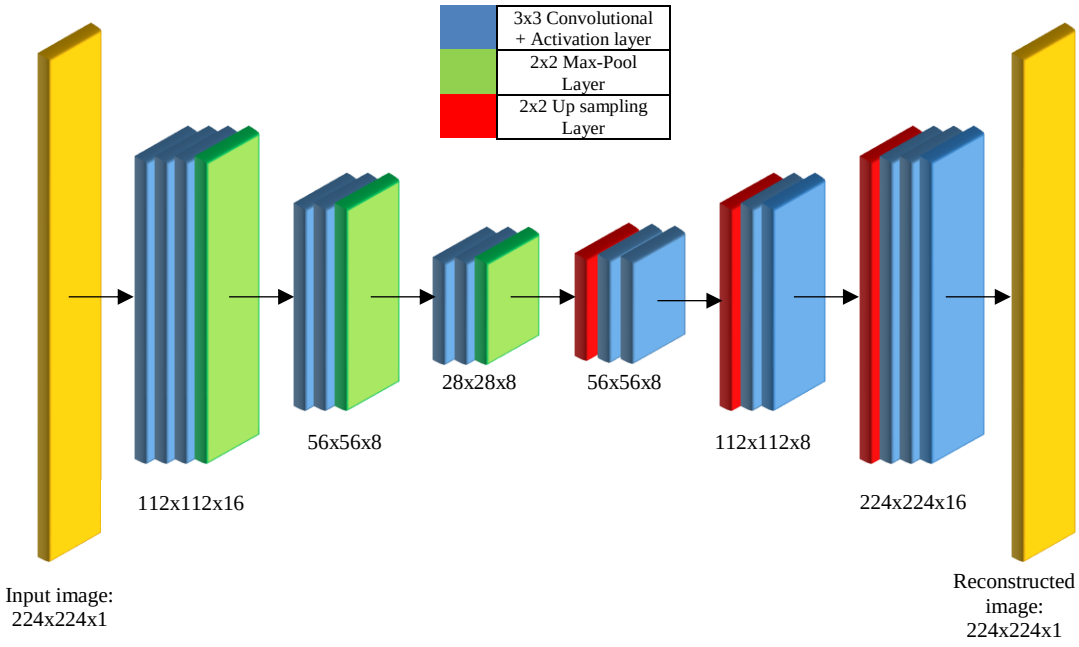


Figure 2.3. Architecture of the one-class encoder. The encoder portion consists of 3x3 convolutional and a 2x2 max-pooling layers. The size of the feature space decreases as we go through the encoder, this prevents the network from learning an identity function. The decoder portion consists of a 3x3 convolutional and a 2x2 up sampling layers

Siamese neural network

Siamese neural networks are a type of supervised learning algorithm that can generalize features from certain image classes even when there are limited examples per class [55]. The network is trained on pairs of images, these images are fed into the same neural network providing two feature vectors. The Euclidean distance between these feature vectors is then computed, each of these

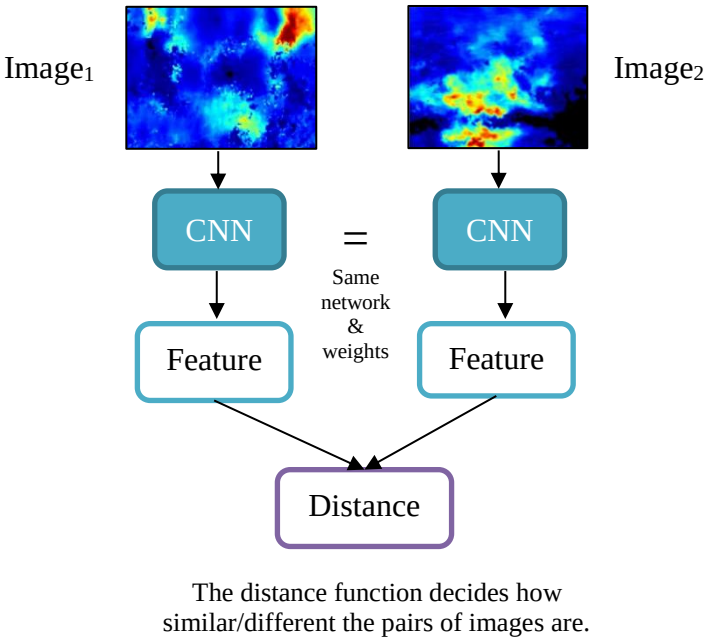


Figure 2.4. Strategy for training a Siamese neural network. If both images belonged to the same class, the distance between the features should be smaller compared to the case when both images belonging to different classes.

feature vectors have 1352 values. If we feed in images belonging to the same class, the network learns weights that minimize the Euclidean distance. On the other hand, if we feed in images belonging to different classes, the network learns weights that maximize the Euclidean distance. This is illustrated in Figure 2.4. The architecture of the neural network used here was the same as the encoder portion of the one class autoencoder illustrated in Figure 2.3. The output of the CNN was flattened such that it takes the shape of a 1-d vector. The Siamese neural network had 2,488 trainable parameters. Like the one-class autoencoder, the number of trainable parameters in the Siamese neural network is much smaller than state of the art convolutional neural networks. We implemented the Siamese neural network using Keras [56].

To evaluate the performance of the Siamese neural network, we first randomly sampled 10 images from the validation set, 5 of these are images of benign nodules and 5 of these are images of malignant nodules. This 10-image set was fixed and served as the comparator set for all testing.

For each test image, we compute the distance of a test image to all 10 images of the comparator set. Next, we compared the distance between the test image and a comparator image of the benign nodule to the distance between the test image and a comparator image of the malignant nodule. This was done in a pairwise manner for all the distances computed between the test image and the 10 comparator images resulting in 25 comparisons per test image. The predicted class of the test image was the class that it gets assigned to the greatest number of times during this pairwise comparison.

2.4.3 *Fine-tuning a neural network*

Fine-tuning a neural network is the process of retraining a neural network for the specific task at

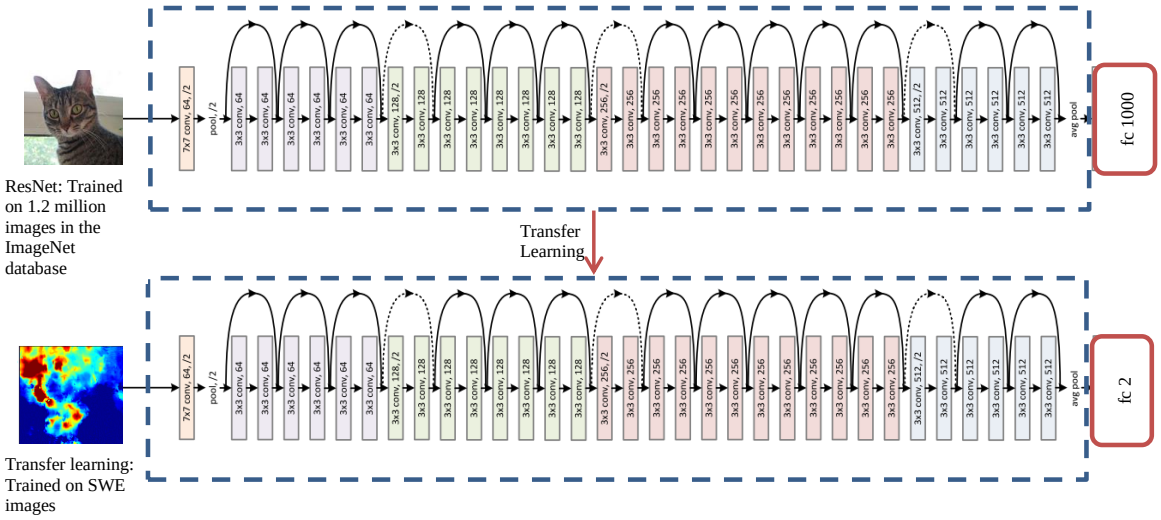


Figure 2.5. Strategy for fine-tuning the ResNet to characterize malignancy in thyroid nodules. In the original training the ResNet model with the ImageNet 2012 data, the final fully connected layer with 1000 outputs was trained. In our application, we replace this layer with a fully connected 2 output layer and train these weights with the ultrasound images.

hand and is a form of transfer learning. In general, these networks have been previously trained on a much larger and sometimes different data set. In our approach, we fine-tuned the ResNet

model, which was previously trained on images in the ILSVRC 2015 data set [42]. The ResNet used a residual learning framework that made it easier to train deeper networks and obtain higher accuracy by increasing its depth. It introduced “skip connections” where by the output of a layer was fed again into the network after skipping a few layers. We used the Keras implementation of the ResNet50 in our analysis [56]. We replaced the last fully connected layer of the original ResNet50 having 1000 outputs with a fully connected layer having 2 outputs. We then added a single drop out layer (as the penultimate layer of the network) and regularization to prevent the network from overfitting. We trained the network by fixing the weights of the initial layers and only updating the weights of the last fully connected layer. The training strategy is illustrated in Figure 2.5. Every instance of the benign nodule was weighted by 0.23 (prevalence of malignant nodule in the training set) and every instance of the malignant nodule was weighted by 0.77 (prevalence of benign nodule in the training set). This way the network was penalized more for misclassifying the image of a malignant nodule compared to misclassifying the image of a benign nodule.

For all the CNN approaches described, we trained one network on the SWE images only and trained another network on the B-mode images only. These were implemented in Keras and details on hyperparameter selections are presented in Table 2.1. The hyperparameters are the type of optimizer used, the learning rate for the optimizer, the drop out ratio for the drop out layer, the L2 regularization parameter and the patience for the early stopping criteria. Patience in this case means the number of sequential epochs during which validation loss does not decrease after which the training process is terminated.

Table 2.1 Algorithmic and hyperparameter selections for our CNN based approaches

Method	Optimizer	Learning rate	Drop out ratio	L2 regularization parameter	Patience – Early stopping criteria
Finely tuning the ResNet50	Adam	1e-5	0.5	0.1	50
One class autoencoder	Adadelta	1e-5	N/A	N/A	10
Siamese neural network	Adam	1e-5	N/A	N/A	20

2.5 RESULTS

In our work, we compared the performance of different machine learning algorithms for the characterization of thyroid nodules. The algorithms were all trained and evaluated on the same training and testing set respectively. For all the algorithms, we have reported metrics such as the accuracy of the network, the specificity at 95% sensitivity and the Area Under the Curve (AUC). Accuracy is defined as the number of times the algorithm correctly identifies the image of the nodule as benign or malignant divided by the total number of images in our testing set. Although accuracy is commonly used to compare the performance of different machine learning algorithms, it might not be the best metric for evaluating the performance of our algorithms. Because our data set is unbalanced classifying all the images in our testing set as benign would give us an accuracy of 0.79. The second metric we considered was specificity at 95% sensitivity because this application warrants operating the classification at a very high sensitivity (95%) to ensure that malignant nodules are not missed, and that people are getting the biopsies that they need. We also report the area under the curve (AUC) computed from the Receiver Operating Curve (ROC). An

AUC of 1 is a classifier that correctly classifies all the instances in our testing set while an AUC of 0.5 describes the performance of a classifier that randomly classifies every instance in our testing set. The classification performance and the ROC curve of a conventional feature extraction plus supervised learning algorithm are described in Table 2.2 and Figure 2.6. The supervised learning algorithms in Table 2.2 have been arranged from highest AUC (Naïve Bayes) to lowest AUC (Decision tree).

Table 2.2 Conventional feature extraction plus supervised learning classification performance

Method	Accuracy	Specificity at 95% sensitivity	AUC
Naïve Bayes	0.80	0.25	0.79
Logistic regression	0.69	0.10	0.64
Support Vector Machine	0.71	0.07	0.62
Decision Tree	0.64	0.05	0.48

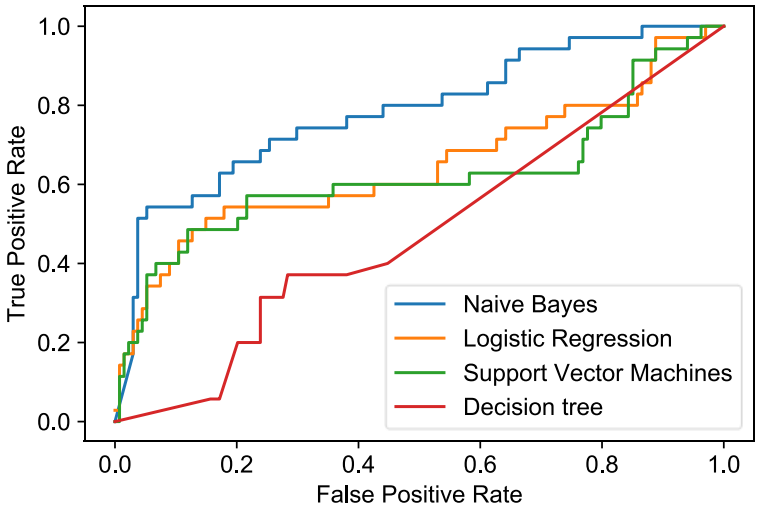


Figure 2.6. ROC curve for different feature extraction plus supervised learning algorithms

The classification performance and the ROC curve for our one-class autoencoder is described in Table 2.3 and Figure 2.7. We obtained the ROC curve by using different thresholds on the MSE's.

The histogram of the MSE's between the original images and the reconstructed images in the test set for each class is presented in Figure 2.8.

Table 2.3 One class auto-encoder classification performance

Method	Accuracy	Specificity at 95% sensitivity	AUC
SWE	0.87	0.40	0.86
B-mode	0.79	0.19	0.45

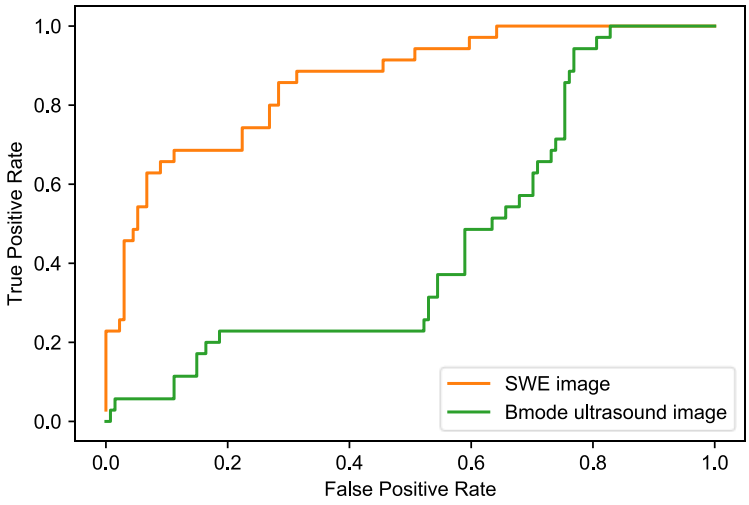
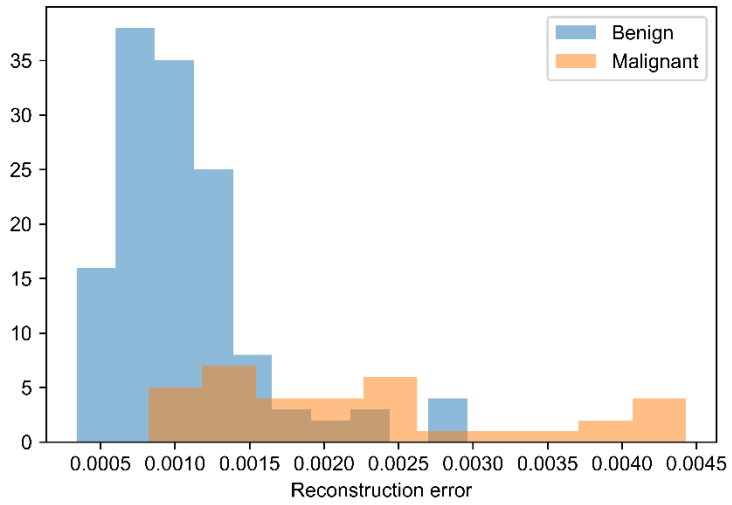
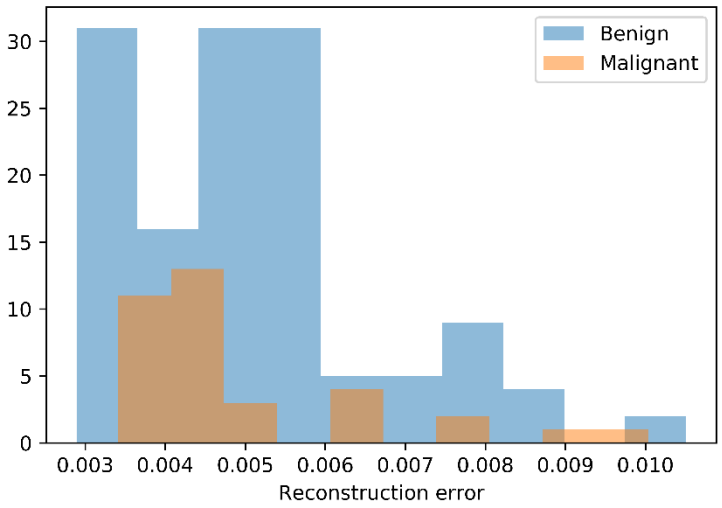


Figure 2.7. ROC curve for a one-class auto encoder



(a)



(b)

Figure 2.8. Histogram of reconstruction errors for a one class autoencoder trained on (a) SWE images only (b) B-mode images only

The classification performance and the ROC curve for a Siamese neural network is described in Table 2.4 and Figure 2.9.

Table 2.4 Siamese neural network classification performance

Method	Accuracy	Specificity at 95% sensitivity	AUC
SWE	0.81	0.41	0.92
B-mode	0.29	0.01	0.35

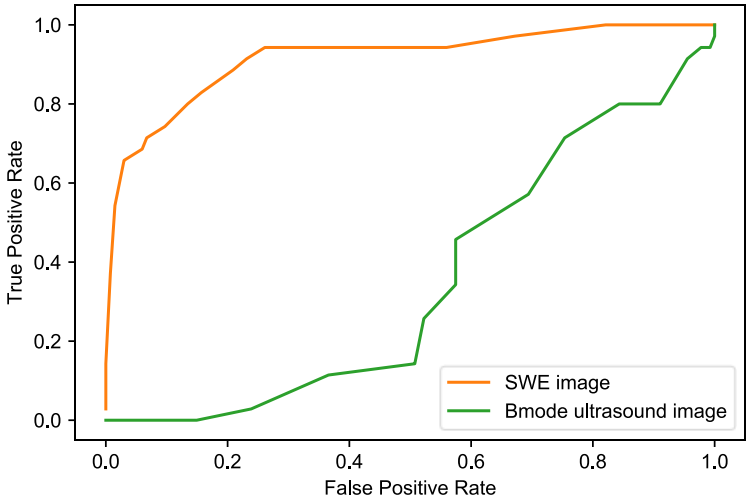


Figure 2.9. ROC curve for a Siamese neural network

The classification performance and the ROC curve for the pre-trained network is presented in Table 2.5 and Figure 2.10. We combined the results of the networks trained on SWE and B-mode images respectively by taking the weighted average of their class scores. We estimated the weights by using a linear regression technique where one input is the score from the network trained on SWE images and the second input is the score from the network trained on B-mode images. The output of the linear regression technique is a score of 0 or 1 where 1 means malignant and 0 means benign.

Table 2.5 ResNet classification performance

Method	Accuracy	Specificity at 95% sensitivity	AUC
SWE & B-mode	0.83	0.55	0.88
SWE	0.85	0.51	0.87
B-mode	0.80	0.10	0.65

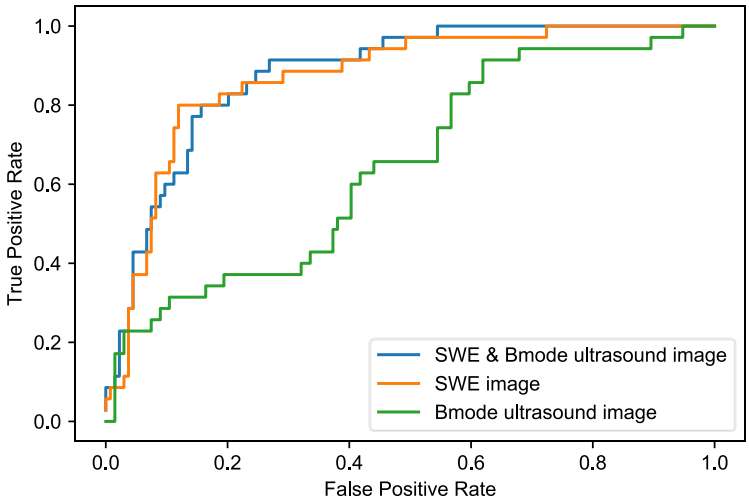


Figure 2.10. ROC curve for a pre-trained network

The AUC of the different classification algorithms are summarized in Table 2.6 along with the number of trainable parameters. The methods have been arranged for smallest to largest number of trainable parameters. The performance of the algorithms is described by its AUC value.

Table 2.6 AUC as a function of trainable parameters for different classification algorithms

Method	Number of trainable parameters	AUC
Naïve Bayes algorithm	34	0.79
Siamese neural network	2488	0.92
Fine-tuning the ResNet	4096	0.87
One-class autoencoder	4385	0.86

2.6 DISCUSSION

We successfully applied a variety of machine learned classification methods to SWE and B-mode ultrasound images of thyroid nodules. For our conventional feature extraction and supervised learning algorithm, the Naïve Bayes classifier gave us the best performance with an AUC of 0.80 as seen in Table 2.2. An advantage of using a conventional machine learning algorithm compared to neural networks is that these could be less prone to overfitting because of the smaller number of trainable parameters. This is more pronounced when the size of the data set is limited as is sometimes the case for medical imaging data sets.

We also trained and evaluated CNNs as a classification method. CNNs differ depending on the architecture and the loss function used [2][14]. The unavailability of a large labelled data set can sometimes make training a CNN for a classification task difficult. For medical imaging tasks, labelling is often an expensive process that might require input from radiologist or pathologist [25]. In our study, we have trained a one-class autoencoder and Siamese neural network. We selected these because of their ability to learn important features from a relatively small data set [51][55]. We also evaluated a fine-tuning approach for our analysis. Previous research has shown that it is possible to finely tune a network that has been previously trained on natural images for a medical imaging task [57]. From Table 2.3, Table 2.4 and Table 2.5 : We obtained comparable performance from our fully trained and fine-tuned CNNs despite the class imbalance and the limited size of our data set. The Siamese neural network fully trained on SWE images had the highest AUC of 0.92. The fine-tuned CNN trained on SWE and B-mode images had the highest specificity of 55% at 95% sensitivity. The one-class auto encoder trained on SWE images had the highest accuracy of 0.87. As seen in Table 2.6, CNN based classification algorithms evaluated in our analysis had a smaller number of trainable parameters compared to training a state

of the art CNNs like the ResNet50 [42]. Having a smaller number of trainable parameters makes these algorithms less prone to overfitting, which is a useful characteristic when the data set is limited. Also note that weighting the images during the training phase helped offset its class imbalance.

Note that we obtained better performance on CNN based approaches that utilized only SWE images over CNN based approaches that utilized only B-mode images. The highest AUC achieved using only B-mode images for a CNN based approach was 0.65. This is much lower than the other approaches that automatically characterized nodules using their B-mode ultrasound images [47]–[49]. One major difference between our approach and published literature is that for published literature, the CNN was trained on segmented B-mode images of the thyroid nodules only. In other words, the previous methods used images that contained only the nodule with limited background tissue. Also, the B-mode images we used were possibly of lower image quality than used by other studies [35]. The B-mode image generated for SWE imaging is generally used to locate the presence of the nodule only and not to assess malignancy of the thyroid nodule. Radiologist assess the thyroid nodule using its B-mode image along its longitudinal and transverse orientations. One drawback of our study is that we treat each of these images independently and we do not combine the information obtained from each of these orientations.

Previous literature suggests that human diagnostic accuracy for the task of thyroid nodule classification using SWE images is on the order of 81.7% [40]. The proposed schemes provide accuracies on par with this ranging from 80 to 87%. Future work would be needed to determine human performance on this same data set. Likewise, additional data is needed to understand the precision of the reported accuracy, specificity and AUC metrics.

2.7 CONCLUSION

In our work we compared the performance of different machine learning approaches for automatic characterization of thyroid nodules using SWE and B-mode images. Among existing CAD systems, a limited number of those utilize SWE images and deep learning in their analysis. All three of our CNN based approaches achieved comparable performance. The performance of our CNN based approaches was similar to the reported diagnostic accuracy of radiologists [40]. Future work will include clinical information such as the TIRAD score in our analysis. We also need to understand the precision of our reported metrics.

Initial results from this project were reported in [A] and final results will be presented in [B]

A. Pereira, Dighe, **Alessio**, "Comparison of machine learned approaches for thyroid nodule characterization from shear wave elastography images," Proc. SPIE 10575, Medical Imaging 2018: Computer-Aided Diagnosis, 105751X (27 February 2018); doi: 10.1117/12.2294572
B. Pereira, Dighe, **Alessio**, "Machine learned approaches for thyroid nodule characterization from shear wave elastography images," in preparation for *IEEE Trans Medical Imaging*, 2018.

Chapter 3. AUTOMATED ANALYSIS OF BREAST LESIONS USING MIP DCE-MRI IMAGES

Magnetic Resonance Imaging (MRI) is an imaging modality that can be used to detect and characterize breast cancer [58]. Among the different MRI techniques, the most popular is dynamic contrast enhanced MRI (DCE-MRI) [59]. This technique is commonly used for screening high-risk patients, monitoring breast cancer during chemotherapy and screening patients with postoperative tissue reconstruction [60]. Currently, contrast enhanced MRI is the most sensitive technique for screening high-risk women [61]. DCE-MRI is reported to have a higher sensitivity of 88% - 100% and a moderate specificity of 68% - 96% [59]. These DCE-MRI images are assessed by radiologists using the American College of Radiology (ACR) MRI Breast Imaging Reporting and Data-System (BI-RADS) [60]. BI-RADS assess malignancy by evaluating morphological features such as shape, margin, lesion type, internal enhancement pattern, and the qualitative assessment of enhancement kinetics of initial uptake [61]. Although this system has helped standardize the diagnosis of breast lesions, studies have shown that there is still variability among radiologists [62]. To facilitate an accurate and quick diagnosis of breast nodules, we evaluate different CNN architectures for identifying breast lesions using Maximum Intensity Projection (MIP) DCE-MRI images.

3.1 LITERATURE REVIEW

Deep learning techniques have been used to characterize breast lesions by utilizing their corresponding DCE-MRI images. Antropova et al. [63] characterized breast lesions by using a pre-trained network in their analysis. The authors fed segmented breast lesions from a post-contrast DCE-MRI slice of the breast to the pre-trained AlexNet[53]. A support vector machine (SVM)

was then trained to classify the lesions as benign or malignant based on the features extracted from the penultimate layer of the AlexNet. The authors achieved an AUC of 0.85 with a database of 551 images. Marrone et al [64], compared different approaches to classify breast lesions using their DCE-MRI images. The authors had a data set of 67 images. The authors segmented the breast lesion from a subtractive DCE-MRI image series, they then used each of these slices in their analysis. They compared three different approaches: fine-tuning the AlexNet, training an SVM on features extracted from the AlexNet, and fully training a network. The feature extraction and fine-tuning method achieved similar performance with an AUC of 0.76 and 0.73. Fully training a network gave the worst performance with an AUC of 0.69. Antropova et al. [65] combined features extracted from a pre-trained CNN and handcrafted radiomics features to analyze malignancy in breast lesions. The authors utilized images of segmented breast lesions from slices of the DCE-MRI volumes at different time points. They achieved an AUC of 0.89 with the DCE-MRI images with a data set of 690 cases. Antropova et al. [66] published another paper on the same data set where they analyzed the performance of MIP images obtained from their subtraction DCE-MRI series and compared it to using just the central slices of the DCE-MRI series at different timepoints. They achieved the highest performance of 0.88 by using MIP images obtained from their subtraction DCE-MRI series.

3.2 Dataset Description

3.2.1 *Image acquisition*

In an IRB approved retrospective study, DCE-MRI images were extracted from the medical PACs system. All exams were performed on either a Philips 3T scanner or a GE 1.5T scanner. DCE-MRI were obtained by injecting contrast into the patients' bloodstream and subsequently capturing a

series of MRI images. Contrast subtracted images were formed from DCE-MRI volumetric data before and after the arrival of contrast using commercial software. This technique can be useful in characterizing malignancy because malignant tumors often have a higher density of blood vessels and higher perfusion marked by higher contrast than surrounding tissue [67]. Although DCE-MRI has improved breast cancer detection and diagnosis, it can be time-consuming to analyze because of the large amounts of data generated [68]. One way to analyze the data effectively was to take the Maximum Intensity Projection (MIP) of the subtracted DCE-MRI volumetric data, thus emphasizing the contrast filled breast lesion [69]. MIP images are obtained by projecting the voxel with the highest contrast value throughout the volume along the superior-inferior axis onto a 2D image [69]. Having the DCE-MRI volumetric data represented by its 2D MIP image can be useful when we want to use a pre-trained CNN like the AlexNet[53] or the ResNet[42] which take in 2D images as input.

3.2.2 *MIP DCE-MRI dataset*

The MIP DCE-MRI data set had approximately 19131 unique data points with pathology confirmed diagnosis. Some patients had multiple studies and each data point in the data set was unique to every breast i.e. a patient with two breasts would have two associated data points, one for each of the breasts. About 19% of the images were malignant while the other images were benign. Each of the data points had a BIRADS score assigned to it. The prevalence and interpretation of the BIRADS score is illustrated in Table 3.1. About 10% of the images did not have a BIRADS score recorded in the database.

Table 3.1. Prevalence and interpretation of BIRADS scores in the MIP DCE-MRI dataset

BIRADS Score	Prevalence (%) in the MIP DCE-MRI dataset	Interpretation
'NA'	10.33	No BIRADS score provided
0	0.13	Need additional imaging or prior examinations
1	33.65	Negative
2	33.19	Benign
3	2.23	Probably Benign
4	4.28	Suspicious
5	0.35	Highly suggestive of malignancy
6	15.85	Known biopsy-proven

(Interpretation based on : http://www.radiologyassistant.nl/en/p53b4082c92130/bi-rads-for-mammography-and-ultrasound-2013.html)

With this data set, we could develop an algorithm to perform a variety of tasks using the MIP DCE-MRI image of the breast. These tasks include predicting malignancy, identifying the presence/absence of a breast lesion and predicting the BIRADS score. In our project, we sought to identify the presence or absence of a breast lesion using its MIP DCE-MRI images. For this task, we divided the images into lesion absent (those with a BIRADs score 1) and lesion present (BIRADs score 3, 4, 5 and 6). Images without BIRADs scores were not included in the analysis. The prevalence of images with a lesion present was 42% while the rest do not have a lesion present.

We performed several preprocessing steps before classifying the images. First, we segmented the MIP DCE-MRI images into left and right breast images. This step included conventional image processing techniques including finding the mid-point between the breasts of each image. Then, each single breast image was cropped to remove empty background and outlier values indicative of imaging artifacts. The segmentation steps have been depicted in Figure 3.1.

The images were then normalized by dividing each of the images with the corresponding maximum intensity of the image. We excluded 182 images in our data set because these images had an artifact, or the patient had a mastectomy. We had a total of 10551 single breast images from 4526 patients for our analysis. These were split into a training (80%) and testing set (20%).

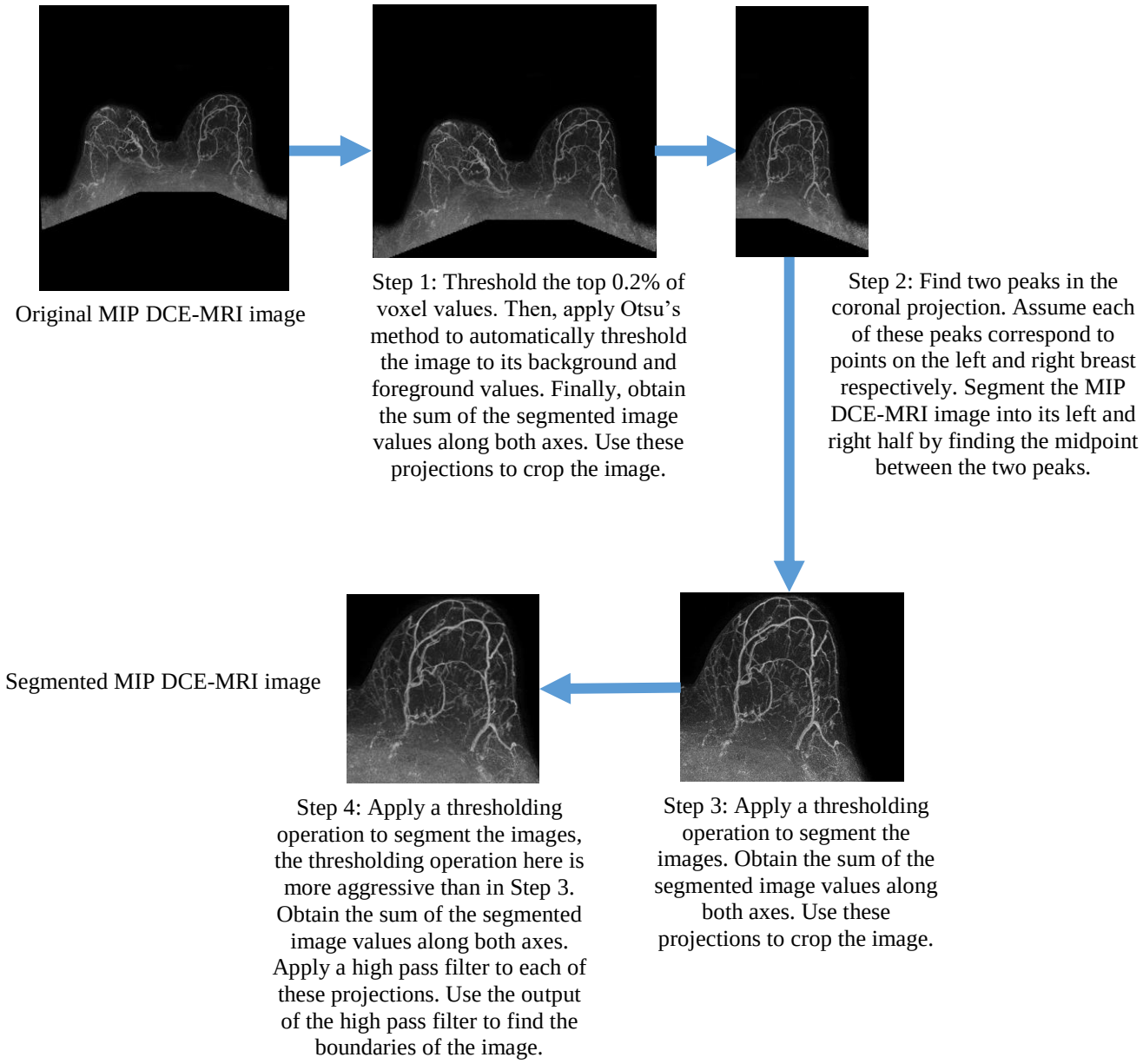


Figure 3.1. Pre-processing steps applied to the original MIP DCE-MRI image

3.3 METHODOLOGY

In our work, we evaluated different deep learning approaches to identify the presence of a lesion using its MIP DCE-MRI image. We pre-trained a ResNet50 [42], fully trained a one-class autoencoder and a Siamese neural network. The theory behind pre-training, the one-class auto encoder and the Siamese neural network has been previously described in Section 2.4.

3.3.1 *Pre-training the ResNet50*

In our approach, we fine-tuned the ResNet50 [42]. We used the Keras implementation of the ResNet in our analysis [56]. We took the output of the 49^{th} layer of the pre-trained network and fed this into a fully connected layer having 32 output neurons followed by a fully connected layer having 2 output neurons. During the training phase, we froze the weights of the first 49 layers of the ResNet and updated the weights of only the last two fully connected layers. We applied an L2 regularization with a factor of 0.05. We used an Adam optimization algorithm with a learning rate of $1e^{-4}$. We implemented an early stopping criterion that stops the training process if the validation loss does not improve after a fixed number of consecutive epochs (20 in this case). We then trained the network again by freezing the first 37 layers and updating the weights of the last 14 layers. We used an Adam optimization algorithm with an initial learning rate of $1e^{-9}$. We chose a much smaller learning rate the second time because we were training a larger set of parameters and we did not want the weights to change too much. We reduced the learning rate by a factor of 0.5 if the validation loss didn't improve after 5 consecutive epochs. We implemented an early stopping criterion again and stopped the training process if the validation loss did not improve after 10 consecutive epochs. We weighted each of the MIP DCE-MRI images with lesion present by 0.81

(prevalence of lesion absent * 1.4) and with no lesion present by 0.19 (prevalence of lesion present *0.6).

3.3.2 *One-Class Auto encoder*

For our analysis we trained a one-class auto encoder to learn the encodings of the images with no lesion present. The architecture of the one-class auto encoder have been previously described in Section 2.4.1 and Figure 2.3 respectively. The architecture was based on the autoencoder implemented using Keras [52]. We trained the network on images of MIP DCE-MRI breast images with no lesion present. The training data had 4320 images and the validation data set had 771 images. We used an early stopping criterion which stopped the training process if the validation loss did not improve after 20 epochs. We used an Adadelta optimization algorithm with a learning rate of $1e^{-4}$.

3.3.3 *Siamese Neural Network*

We also trained a Siamese neural network. The theory and architecture of the Siamese neural network have been previously described in Section 2.4.1 and Figure 2.4 respectively. The architecture of the CNN used is the same as the architecture of the encoder portion of the auto encoder. The Siamese neural network was trained on 29736 pairs of images. Half of these pairs had images belonging to the same class and the other half of these pairs had images belonging to different class. The validation set had 5244 pairs of images. Like we did for the pre-trained network and the one-class auto encoder, we used an early stopping criterion which stopped the training process if the validation loss did not improve after 20 epochs. We used an Adam optimizer with a learning rate of $1e^{-5}$.

3.4 RESULTS

In our work, we compared the performance of different classification algorithms for detecting the presence of a breast lesion given its corresponding MIP DCE-MRI image. We reported metrics such as accuracy, AUC and the specificity at 95% sensitivity. The background for all three metrics has been described in Section 2.5. The classification performance of the three neural networks used in our analysis are reported in Table 3.2.We arranged the algorithms from highest AUC value (Fine-tuning) to lowest AUC value (Siamese neural network).

Table 3.2 Classification performance of deep learning-based classification algorithms

Method	Accuracy	Specificity at 95% sensitivity	AUC
Fine-tuning	0.69	0.19	0.73
One-class Autoencoder	0.57	0.10	0.54
Siamese neural network	0.56	0.05	0.5

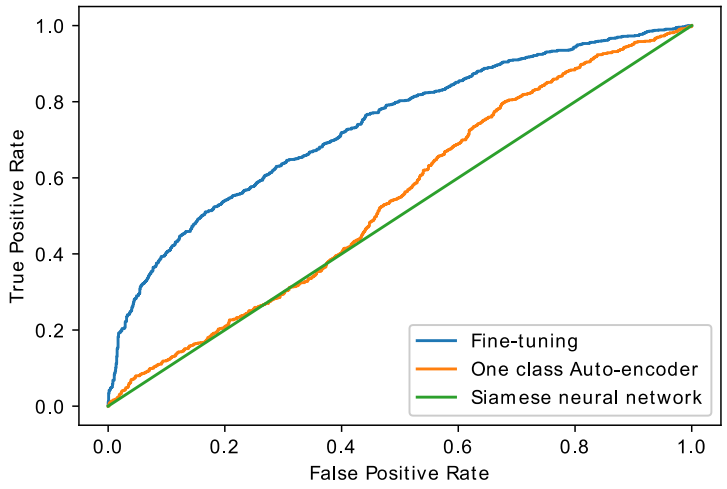


Figure 3.2. ROC curve for different deep learning-based classification algorithms

The histogram of classification scores generated from the pre-trained network for all the test images have been depicted in Figure 3.3. When an image gets a score greater than 0.5, it is classified as having a lesion present while an image with a score of less than 0.5 is classified as having no lesion present/lesion absent. The closer the score is to 1, the more confident the algorithm is that a lesion is present while the closer the score is to 0, the more confident the algorithm is that no lesion is present.

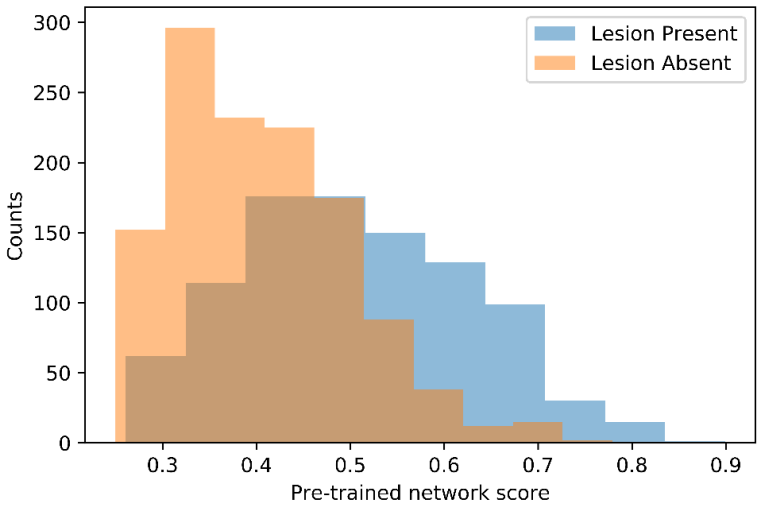


Figure 3.3. Histogram of classification scores from the pre-trained network

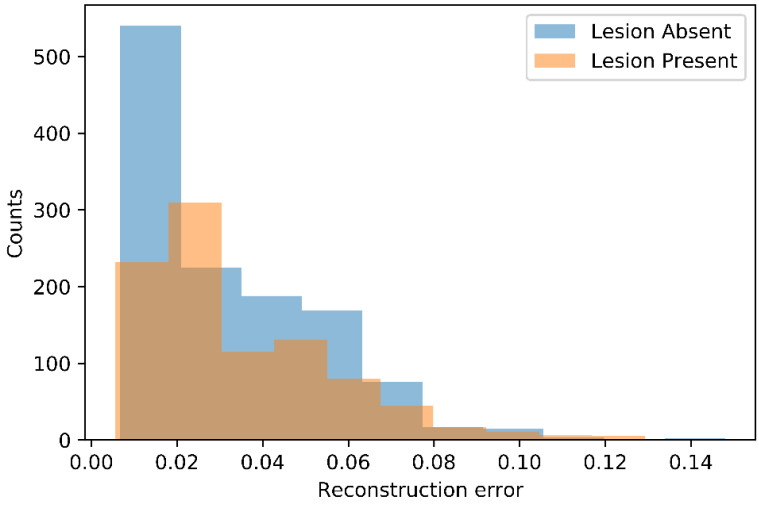


Figure 3.4. Histogram of reconstruction errors from the one class auto encoder

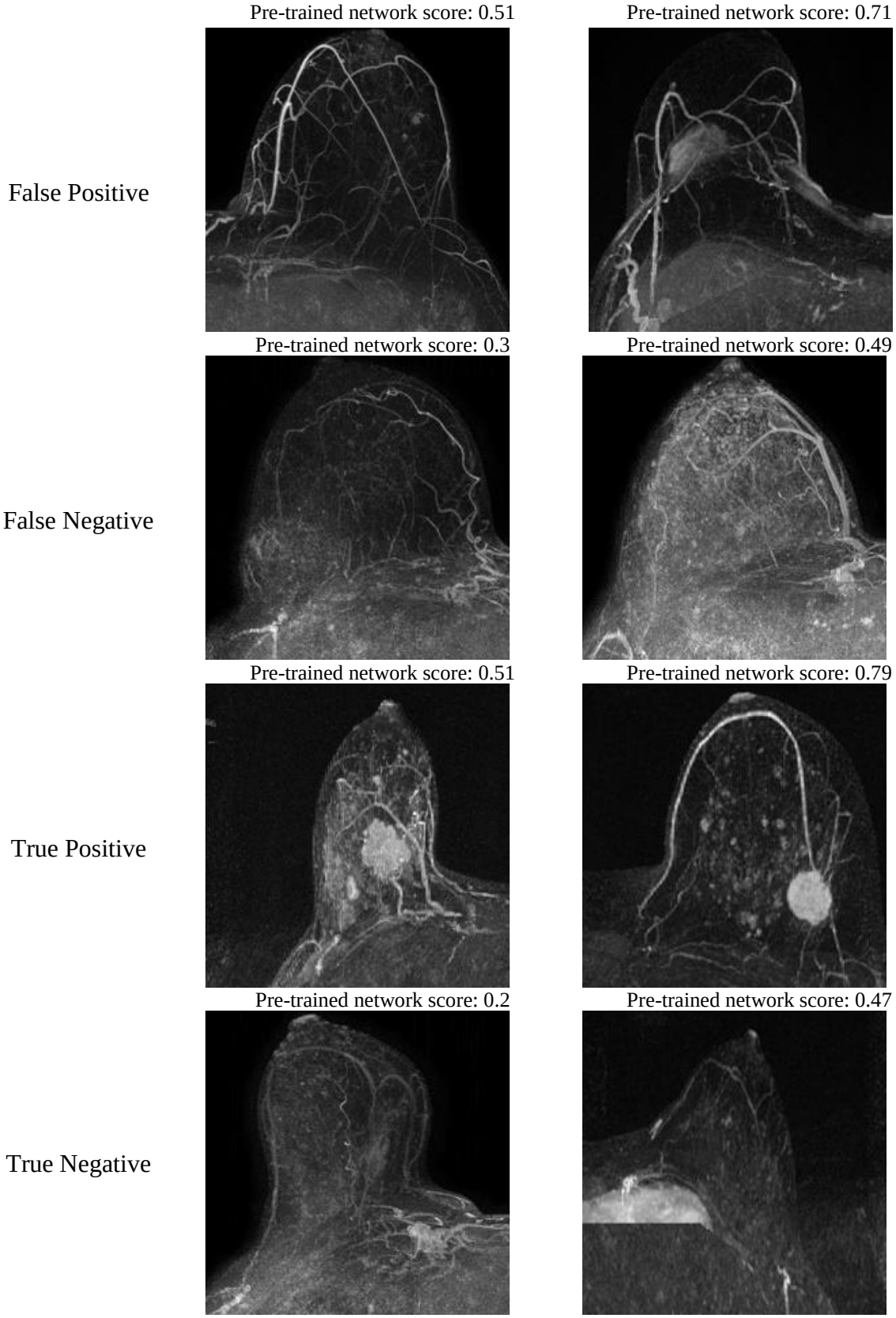


Figure 3.5. Example MIP DCE-MRI test images with their corresponding classification score

The histogram of reconstruction errors/MSE between the original image and the image generated from the auto encoder is depicted in Figure 3.4. Figure 3.5 are example of test images that have been misclassified and correctly classified by the pre-trained network along with their corresponding classification score. Images that are false positive are classified as having a lesion present when there is no lesion present. Images that are false negative are classified as having no lesion present when there is a lesion present. Images that are true positive are correctly classified as having a lesion present and images that are true negative are correctly classified as having a lesion absent.

Table 3.3 Number of learnable parameters

Method	AUC	Number of learnable parameters
Fine-tuning	0.73	3,486,306
One-class autoencoder	0.54	4385
Siamese neural network	0.5	2488

3.5 DISCUSSION

We applied different deep learning-based classification algorithms to analyze MIP DCE-MRI images of the breast. The goal of our analysis was to develop a classification algorithm that would identify if a MIP DCE-MRI breast image had a lesion present. We obtained the best performance using the pre-trained ResNet with an AUC value of 0.73 followed by the auto encoder with an AUC value of 0.54 and a Siamese neural network with an AUC value of 0.5. The Siamese neural network (AUC = 0.5) had the worst performance. It performed as well as a trivial classifier which classified all the test images as the majority class i.e. the lesion absent class. This might be due to the smaller number of trainable parameters compared to the one-class auto encoder and the pre-

trained network. The one class auto encoder (AUC = 0.54) did only slightly better than the Siamese neural network, this might also be due to the smaller number of trainable parameters compared to the pre-trained network. Among the different classification algorithms considered, the pre-trained network gave us the best performance, this might be because the weights and architecture have been previously verified albeit for a different data set. The pre-trained network also had the highest number of trainable parameters compared to the one class autoencoder and the Siamese neural network. The results from the pre-trained network showed that among the misclassified images, there were a greater number of false negatives than there were false positive. This was despite the cost of misclassifying an image with lesion present (cost = 0.81) being substantially higher compared to the cost of misclassifying an image with no lesion present (cost = 0.19).

Similar work has been performed previously. Antropova et al. achieved an AUC of 0.88 for detecting malignancy in the breast using MIP DCE-MRI images. Their classification task of malignancy versus benign was different from ours (lesion present versus absent). In addition, their study had substantially smaller data set of only 690 images [66]. For comparison, in our analysis we had 10551 MIP DCE-MRI images of the breast. One important difference is that Antropova et. al aimed to classify known lesions as benign or malignant, rather than to identify the presence of lesions as in our study and trained the network on segmented images that only contained the breast lesion instead of training on the entire MIP DCE-MRI image of the breast, which may reduce the complexity of the task. Note that they achieved an AUC of 0.88 using a pre-trained network. Recent studies that have used deep learning algorithms and DCE-MRI images for detecting malignancy in the breast have all utilized segmented images that only contained the lesion [63]–[66]. The network that had the most number of trainable parameters also gave us the best performance indicating that in the future we could explore more complex algorithms.

3.6 CONCLUSION

In our work, we compared the performance of different deep learning-based classification algorithms for identifying the presence of a breast lesion given its corresponding MIP DCE-MRI image. We obtained the best performance (AUC = 0.73) by pre-training the ResNet [42]. This was slightly lower than the reported AUC of 0.76 to 0.88 by other groups [63]–[66], although it should be noted that tasks for detection of a lesion versus detection of malignancy varied amongst these studies. Our work appears to be among the first to utilize whole breast images as opposed to images requiring an expert to pre-segment individual lesions. Future efforts with this data set for classification would seek to identify malignant versus benign disease and potentially classify each breast according to its highest risk BIRADS score.

Chapter 4. DEEP LEARNING BASED UPSAMPLING SCHEMES FOR ULTRASOUND IMAGES

Ultrasound is a commonly used medical imaging modality. It is a portable, low-cost, and non-ionizing device that has been used for diagnostic purposes in fields such as obstetrics and cardiology [70]. Diagnostic accuracy of any medical imaging modality is related to its image quality which in turn is related to the spatial resolution of the image [71]. The spatial resolution of an ultrasound image is normally dictated by the features of the transducer and the ultrasound beam that it creates [72]. Besides optimizing the features of the transducer, we can also improve the resolution of the images through post-processing. A common way to improve the resolution of an image is through interpolation-based techniques such as bi-cubic interpolation [73] and Lanczos resampling [74]. Recently, deep learning based upsampling schemes have been used to upsample high resolution (HR) images from their corresponding low resolution (LR) image [75]. Upsampling is a well-researched area in computer vision [75] which is why we want to evaluate the practicality of using such algorithms for upsampling medical images where there is a risk of misdiagnosis. In this chapter, we train and evaluate a deep learning-based algorithm for upsampling ultrasound images. We also report the artifacts generated.

4.1 LITERATURE REVIEW

Most deep learning based upsampling schemes utilize CNNs as their basic framework which can learn representations from images directly. This is useful for upsampling because mapping a HR image from its LR image is an ill-posed problem and a LR image can be mapped to different possible HR images [75]. In this section, we discuss different deep learning based upsampling schemes. Super Resolution Convolutional Neural Network (SRCNN) was one of the first deep

learning based upsampling schemes introduced [76]. It is a three-layer CNN that learns the mapping from a LR image to a HR image. The input to the image is a bicubically interpolated LR image. The loss function for optimizing the SRCNN is the mean squared error (MSE) [76]. More recent deep learning based upsampling schemes take in a LR image directly without interpolation – they utilize more complex architectures and different loss functions that better reflect the qualities of a HR image. Networks such as FSRCNN [77] have deconvolutional layers that increase the resolution of the image gradually as it passes through the network. This approach allows the CNN to learn a better approximation of the HR image compared to the SRCNN. In various other applications, it has been shown that deeper architectures tend to have better performance [75]. The VDSR network is a 20 layer network that learns the residual image between the bicubic LR image and the HR image [78]. The authors utilize a higher initial learning rate and gradient clipping to train the deeper network. SRResNet is another deep network that uses residual layers to learn the mapping between a HR and a LR image [79]. The network utilizes 16 residual units. Most of these networks minimize the MSE as its objective function. The resulting images have higher SNR compared to their corresponding bi-cubic interpolated images, however, these images can lack the high frequency details associated with HR images.

The SRGAN utilizes a Generative Adversarial Network (GAN) architecture for upsampling [79]. In general, a GAN is composed of a generator part and a discriminator part. The SRGAN relies on an adversarial loss and a content loss instead of just an MSE loss. The generator part generates upsampled images from its LR image while the discriminator differentiates between the upsampled image and its corresponding HR image. The learning process is complete once the discriminator is unable to differentiate between the upsampled image and the HR image. This training scheme

allows the generator to generate images that have high frequency information making it perceptually more appealing compared to some of the other deep learning based upsampling schemes.

4.2 Dataset Description

We trained the SRGAN on B-mode ultrasound images of the breast. These ultrasound images were obtained from DICOM cine loops provided by Philips Healthcare. We decimated the cine loop temporally in such a way that selected images were not highly correlated with each other. We also cropped the images such that it only contained B-mode information and the aspect ratio was kept constant. We trained the SRGAN on dataset containing 1921 training images and 481 validation images. The training and validation data set were obtained from two sites. We evaluated the performance of the SRGAN on a dataset containing 333 images, these were obtained from a third independent site.

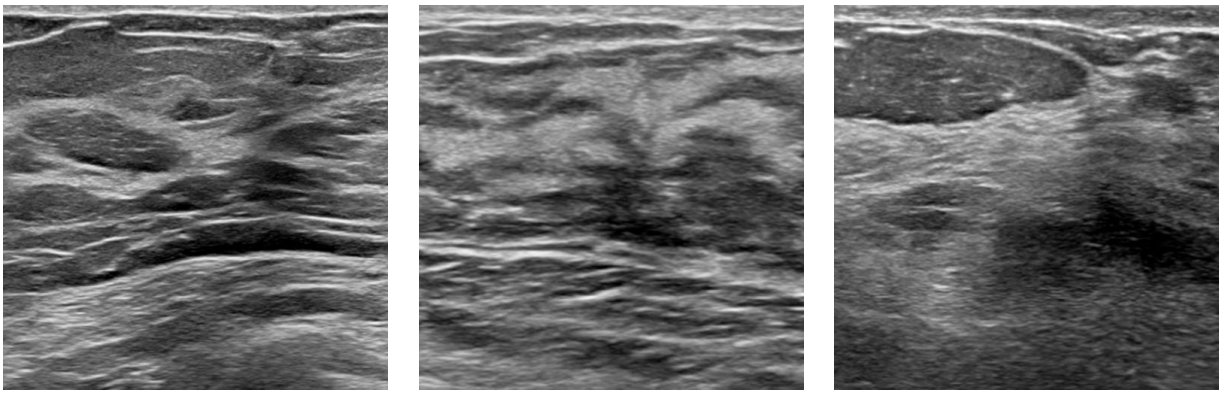


Figure 4.1. Example ultrasound B-mode images used for training the SRGAN

The HR images were 512x512 and the LR images were 128x128, these were obtained by decimating their corresponding HR image by a factor of 4. Example HR images used for training the SRGAN are depicted in Figure 4.1.

4.3 METHODOLOGY

In our work, we train the SRGAN to upsample ultrasound images. The network was originally designed for use on natural images. The performance of the network is illustrated in Figure 4.2. Example of upsampled image generated using the SRGAN. that the upsampled image resemble

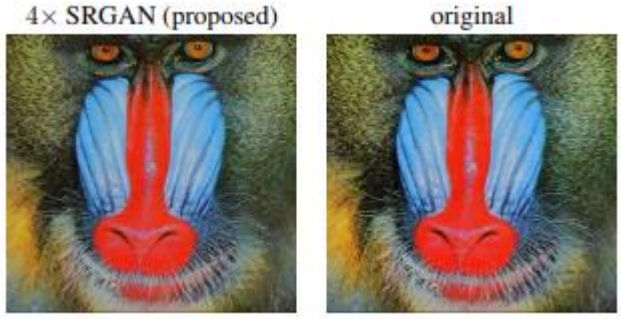


Figure 4.2. Example of upsampled image generated using the SRGAN.[79]
Source: https://arxiv.org/pdf/1609.04802.pdf

the original HR image.

We trained the SRGAN using the publicly available code on github [80]. The SRGAN was trained by randomly cropping image patches of size 256x256 from the HR image. These patches were then down sampled by a factor of 4 to obtain LR patches of size 64x64. This patch-based approach has been used to train various deep learning based upsampling schemes [75]. It allows us to artificially expand our data set and speed up the training process. We first trained the generator portion of our SRGAN only. Once the training and validation loss has decreased for the generator, we trained both the generator and discriminator simultaneously. We used the ADAM optimizer with an initial learning rate of 1e-4.

We computed the Mean Square Error (MSE), Peak Signal to Noise Ratio (PSNR) and Natural Image Quality Evaluator (NIQE) for the images upsampled using bicubic interpolation and the SRGAN. MSE measures the mean of the squares of the error between the upsampled

image and the HR image. PSNR measures the ratio of the maximum power of the signal to the power of the noise in the signal. Both the MSE and the PSNR are commonly used to evaluate upsampling algorithms. However, neither the MSE nor the PSNR measures the ability of the upsampling algorithm to capture perceptually relevant characteristics [79] i.e. lower MSE and higher PSNR is not necessarily related to better image quality. We also compare the NIQE scores for each of the upsampling schemes. We first trained the NIQE model on the HR test images. In order to evaluate an upsampled test image, the NIQE measures the distance between features of the upsampled image to the features obtained from the HR test images. These features are modeled as Gaussian distribution. The NIQE has shown to be related to human preference for high resolution images with lower NIQE being related to better image quality [81]. It is still, however, not a perfect predictor of image quality.

Initial results were promising but a few of the images had a checkerboard artifact present. These artifacts had been previously seen in images generated by other neural networks [82]. Odena et al. have demonstrated that replacing the deconvolution layer with a "resize-convolution" layer can remove the artifacts in the generated images [82]. The resize-convolution layer is an interpolation operation followed by a convolution layer. The authors found that a nearest neighbor interpolation worked best, but this could differ depending on the task. In our work, we have replaced the two deconvolution layers in the SRGAN with a nearest neighbor interpolation and a convolution layer having 64 filters with kernel size 3x3. Each of the nearest neighbor interpolation operation upsamples the image by a factor of 2. An additional artifact that was seen in the original SRGAN was a shift in pixel intensities. The upsampled image appeared to have lower image pixel values compared to the original HR image. We addressed this artifact

by applying a histogram equalization algorithm between the upsampled image and the bicubic interpolated versions of the LR image.

4.4 RESULTS

We evaluated the performance of different versions of the SRGAN on the testing set. LR images of size 128x128 were given as input to the network. The SRGAN upsamples the LR image by a factor of 4 and outputs an image of size 512x512. We compare the image upsampled using both the modified SRGAN and the original SRGAN with the image upsampled using bi-cubic interpolation. The original SRGAN has the same architecture as the network described in [79] while for the modified SRGAN we replace the deconvolution layers of the original SRGAN with 'resize-convolutional' layers. We also evaluate the performance of the histogram equalization algorithm on the test images. Our test data set had about 333 images. The MSE, PSNR and NIQE of the different upsampling schemes are reported in Table 4.1.

Table 4.1. Results of upsampling schemes

	MSE (mean, standard deviation)	PSNR (mean, standard deviation)	NIQE (mean, standard deviation)
Bicubic interpolation	87.8, 33.8061	28.9, 1.5	14.7, 1
SRGAN	149.5, 95.4	27, 2.2	7, 1.2
Modified SRGAN	119.9, 39.2	27.5, 1.4	6.4, 0.9
Modified SRGAN with histogram equalization	126.3, 37	27.3, 1.3	6.4, 0.9

Figure 4.3, Figure 4.4 and Figure 4.5 illustrates the MSE, the PSNR and the NIQE of each of our upsampling schemes for the individual test images. We also plot the ratio of the MSE, the PSNR and the NIQE of each of our upsampling schemes to the corresponding metric computed for the bicubic interpolated image. For MSE and PSNR, ratios of less than 1 indicate better performance

than bicubic interpolation while for NIQE ratios of greater than 1 indicate better performance than bicubic interpolation.

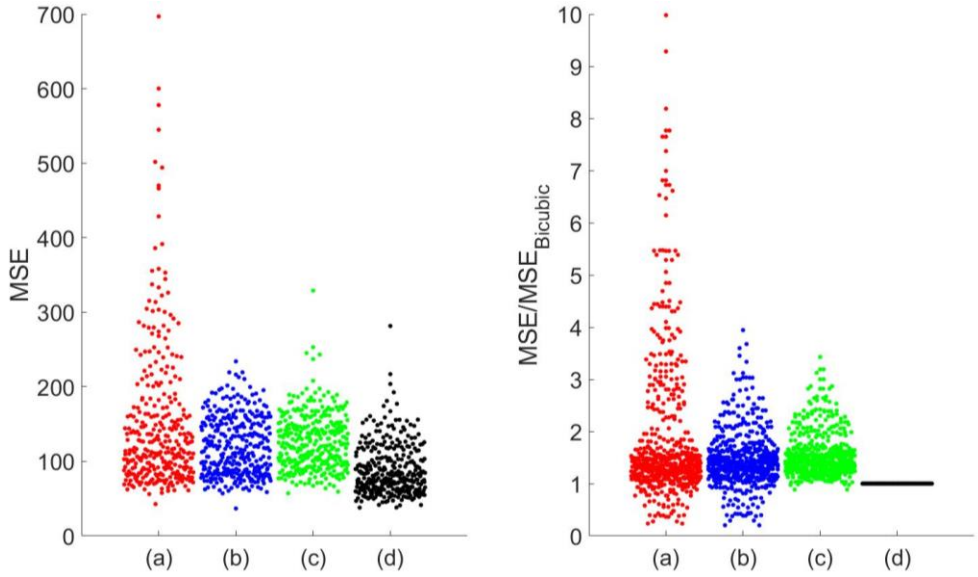


Figure 4.3. Left: MSE of test images upsampled using different algorithms. Right: Ratio of the MSE of test images upsampled using different algorithms to the MSE of the bicubic interpolated image. Upsampling algorithms considered are (a) SRGAN (b) modified SRGAN (c) modified SRGAN with histogram equalization and (d) bicubic interpolation. Lower MSE implies better performance.

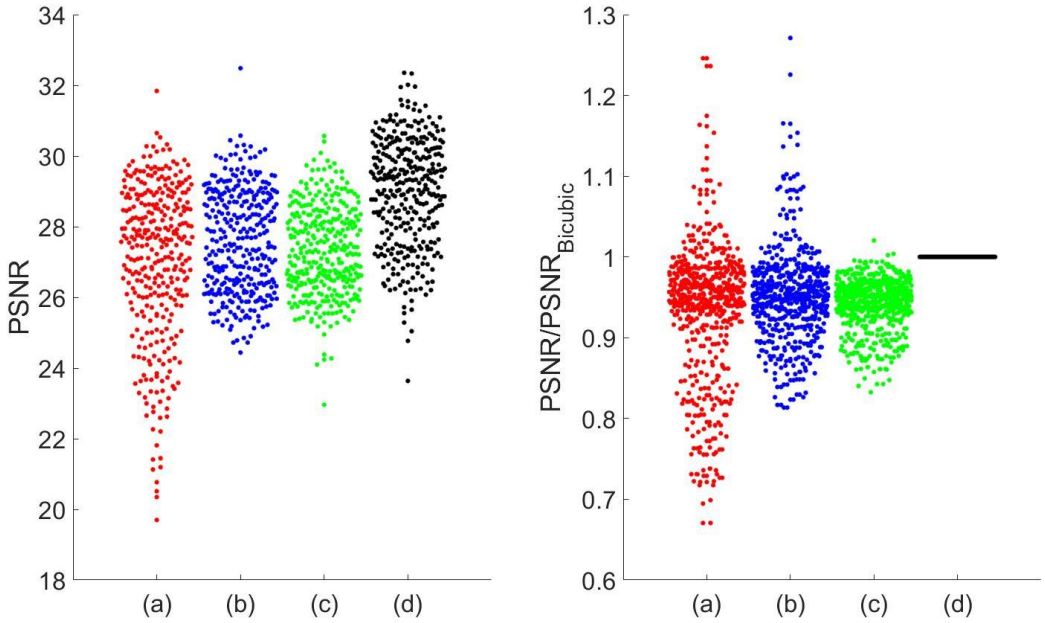


Figure 4.4. Left: PSNR of test images upsampled using different algorithms. Right: Ratio of the PSNR of test images upsampled using different algorithms to the PSNR of the bicubic interpolated image. Upsampling algorithms considered are (a) SRGAN (b) modified SRGAN (c) modified SRGAN with histogram equalization and (d) bicubic interpolation. Higher PSNR implies better performance.

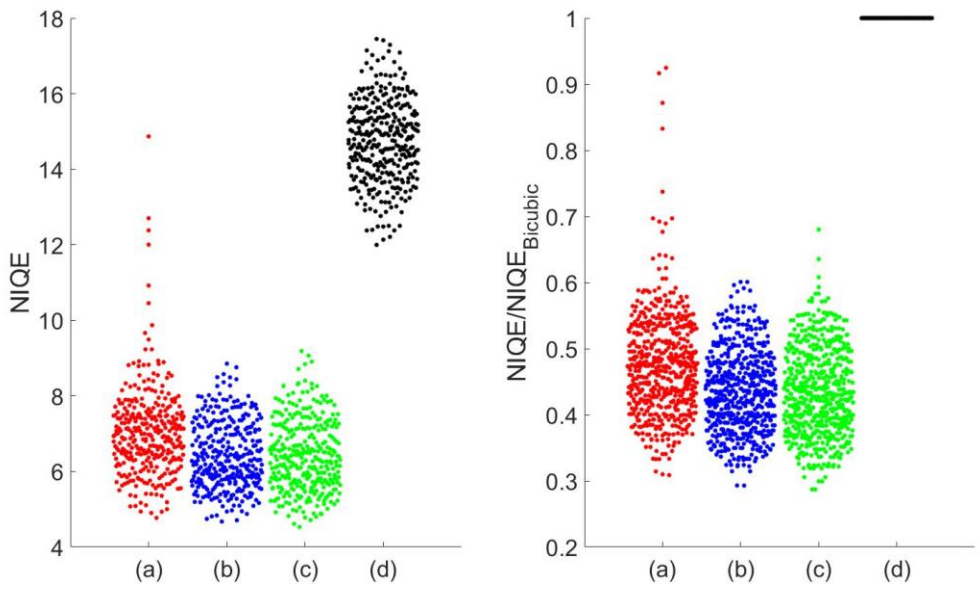
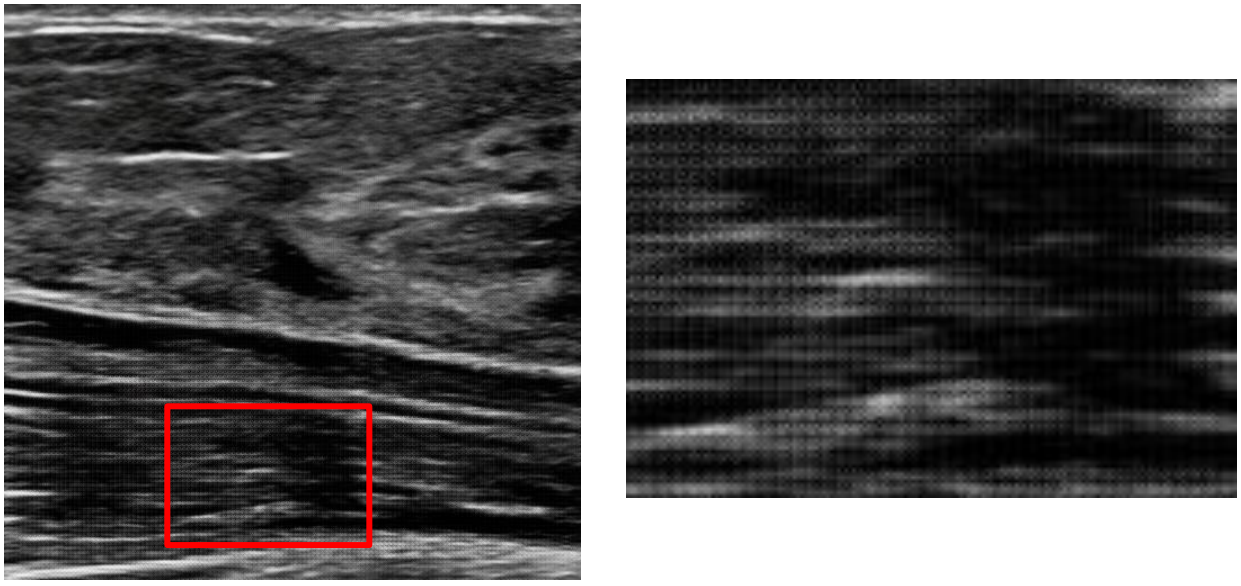
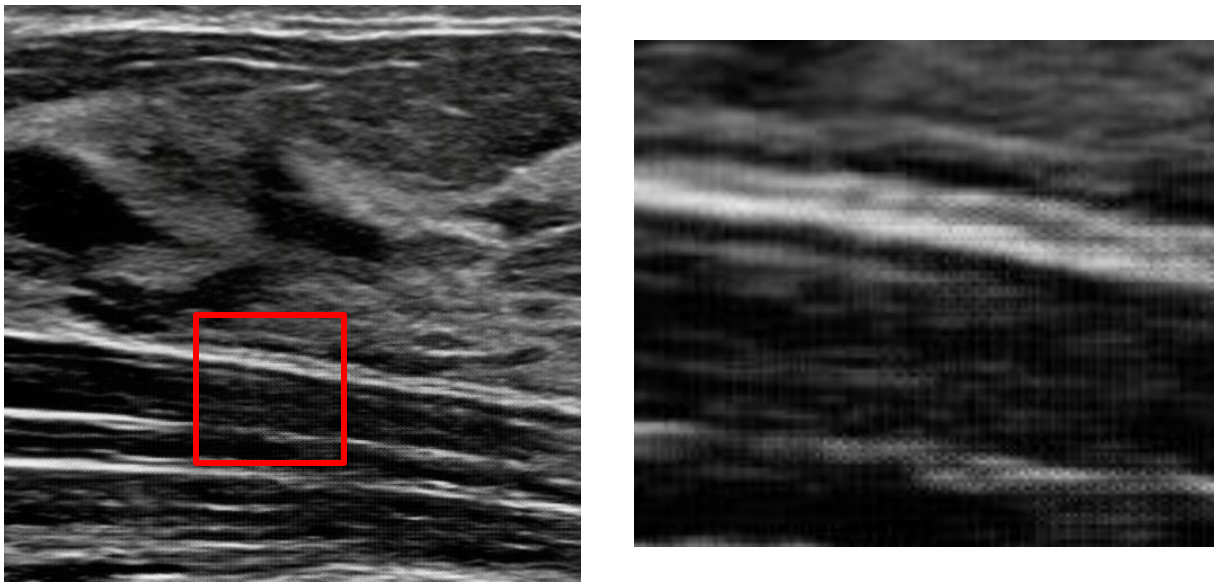


Figure 4.5. Left: NIQE of test images upsampled using different algorithms. Right: Ratio of the NIQE of test images upsampled using different algorithms to the NIQE of the bicubic interpolated image. Upsampling algorithms considered are (a) SRGAN (b) modified SRGAN (c) modified SRGAN with histogram equalization and (d) bicubic interpolation. Lower NIQE implies better performance.



1.a. 1.b.



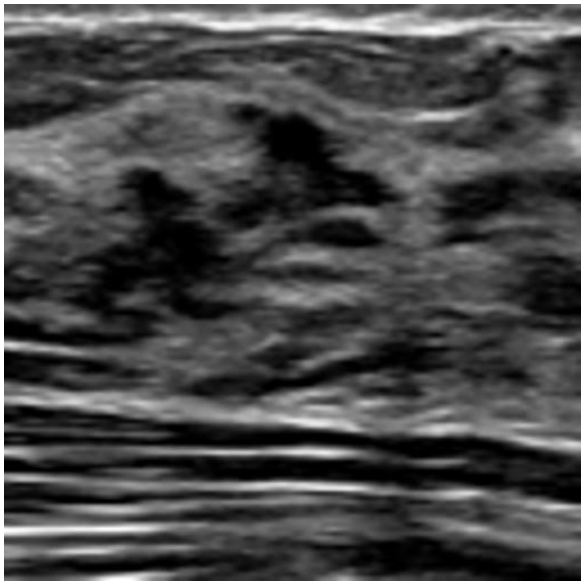
2.a. 2.b.

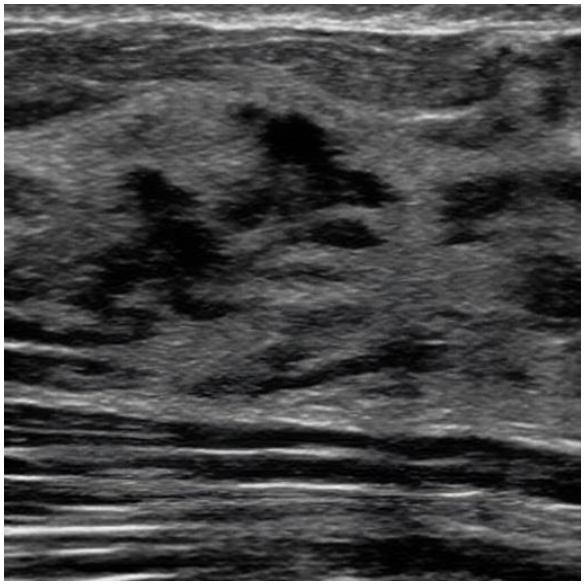
Figure 4.6. Example images upsampled using SRGAN with checkerboard artifacts present. 1.b. and 2.b. highlight the checkerboard artifacts present in 1.a. and 2.a. respectively.

Figure 4.6 illustrates the checkerboard artifact present when training on the original SRGAN architecture.

(a) Bicubic interpolation (MSE: 122.4, NIQE: 16)

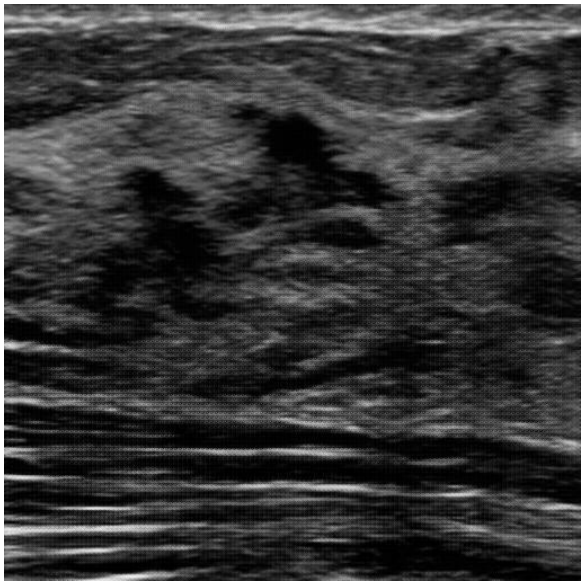
(b) Modified SRGAN (MSE: 197.8, NIQE: 7.71)





(c) SRGAN (MSE: 501.5, NIQE: 14.8)

(d) Original HR image



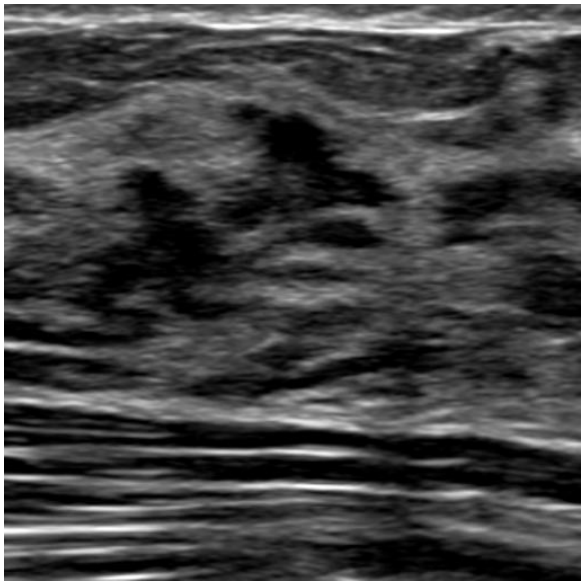
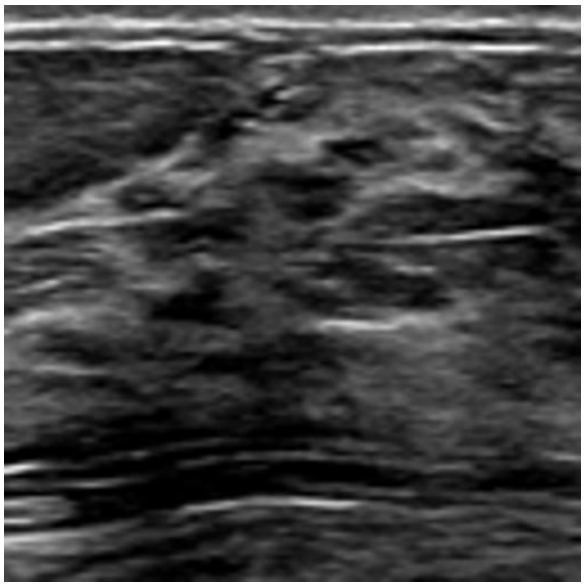
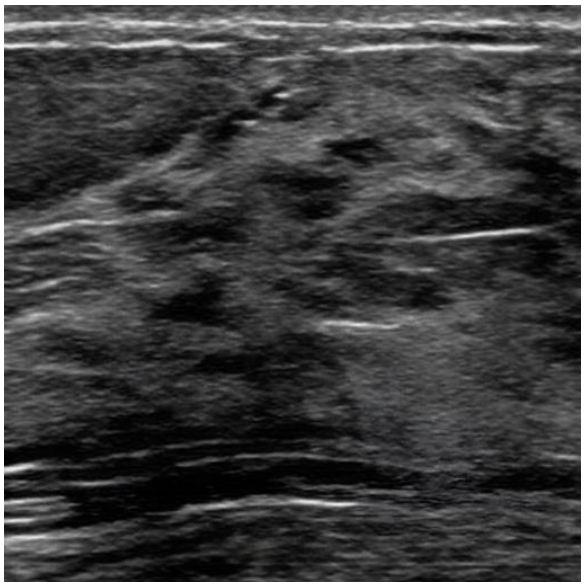


Figure 4.7. Examples of upsampled test images. The images were upsampled using (a) Bicubic interpolation (b) SRGAN (c) Modified SRGAN (b) Original HR image. Corresponding MSE and PSNR are shown in the bracket.

(a) Bicubic interpolation (MSE: 86.5, NIQE: 14.6)

(b) Modified SRGAN (MSE: 159.2, NIQE: 8.76)





(c) SRGAN (MSE: 239.8, NIQE: 12.70)

(d) Original HR image

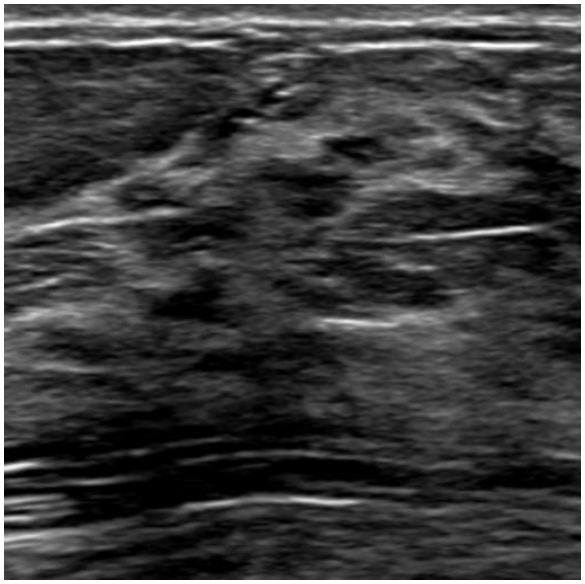
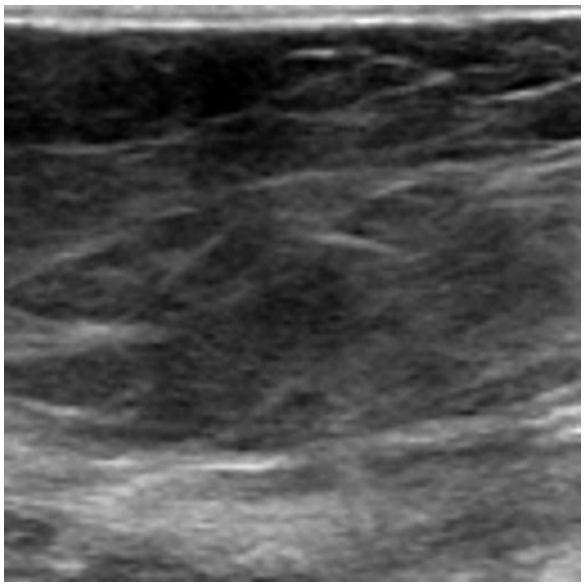
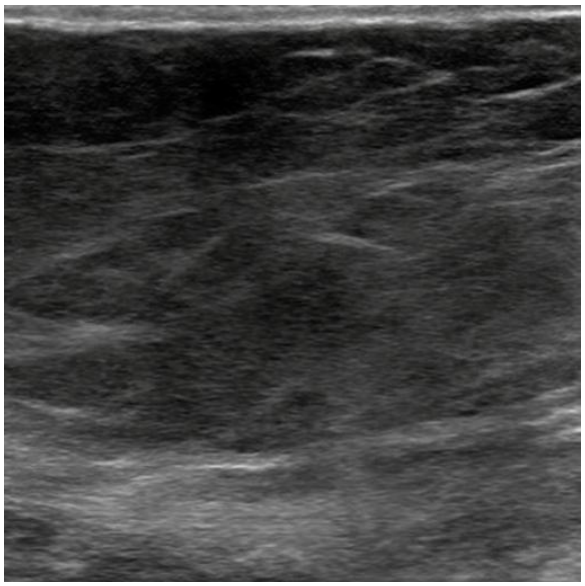


Figure 4.8. Examples of upsampled test images. The images were upsampled using (a) Bicubic interpolation (b) SRGAN (c) Modified SRGAN (b) Original HR image. Corresponding MSE and PSNR are shown in the bracket

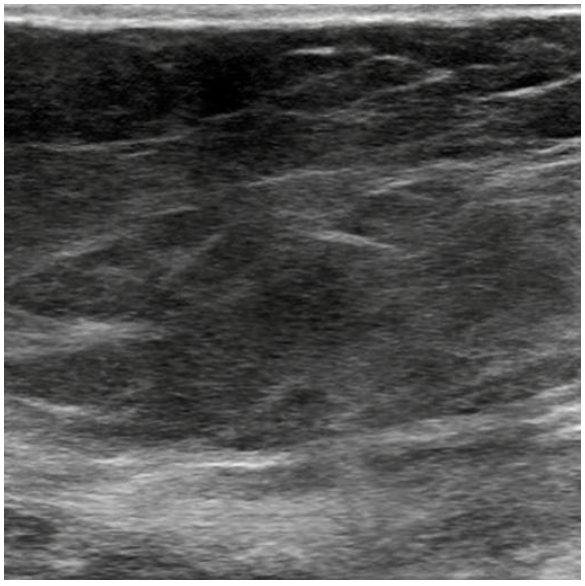
(a) Bicubic interpolation (MSE: 281.1, NIQE: 15.3)



(b) Modified SRGAN (MSE: 82.5, NIQE: 8.5)



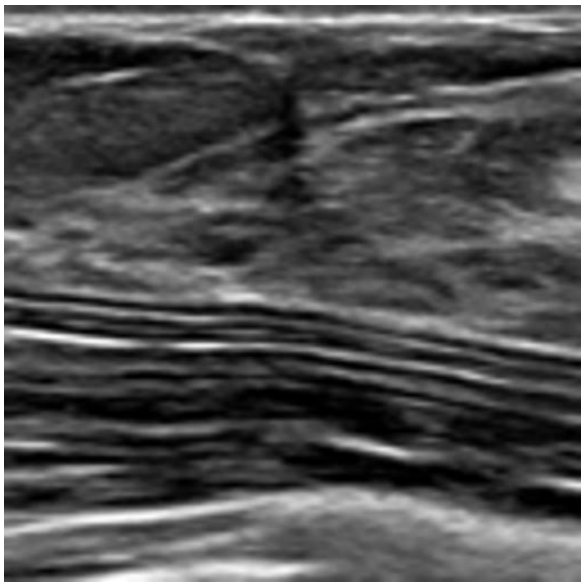
(c) Modified SRGAN + Histogram equalization (MSE: 328.9, NIQE: 9.1)



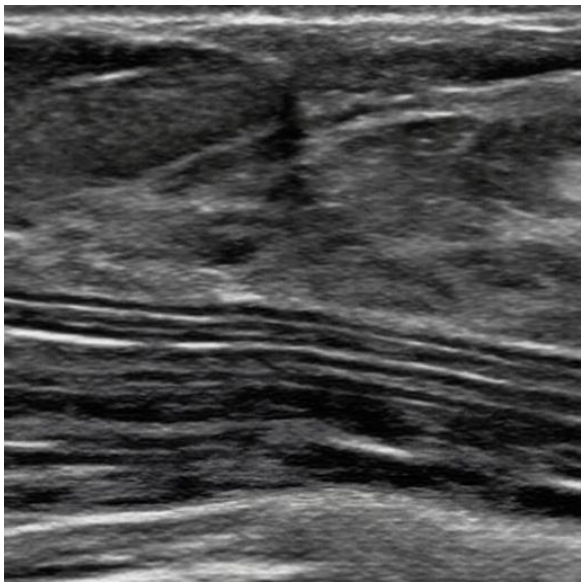
(d) HR image

Figure 4.9. Examples of upsampled test images. The images were upsampled using (a)Bicubic interpolation (b) Modified SRGAN (c) Modified SRGAN with histogram equalization (d) Original HR image

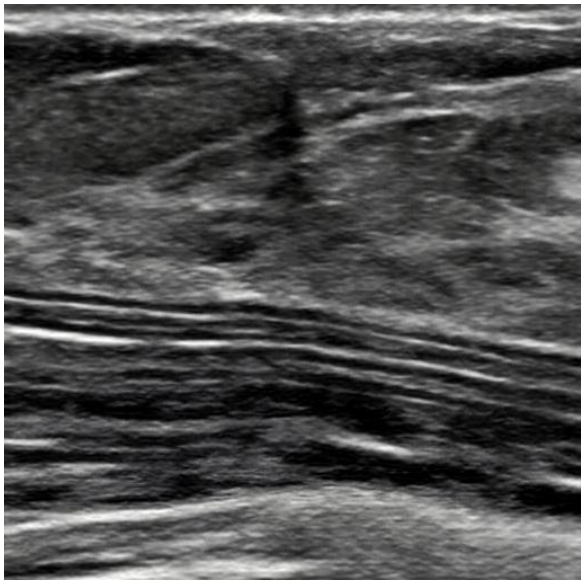
(a) Bicubic interpolation (MSE: 55.4, NIQE: 15.3)



(b) Modified SRGAN (MSE: 184.6, NIQE: 5.9)



(c) Modified SRGAN + Histogram equalization (MSE: 147.8, NIQE: 6.2)



(d) HR image

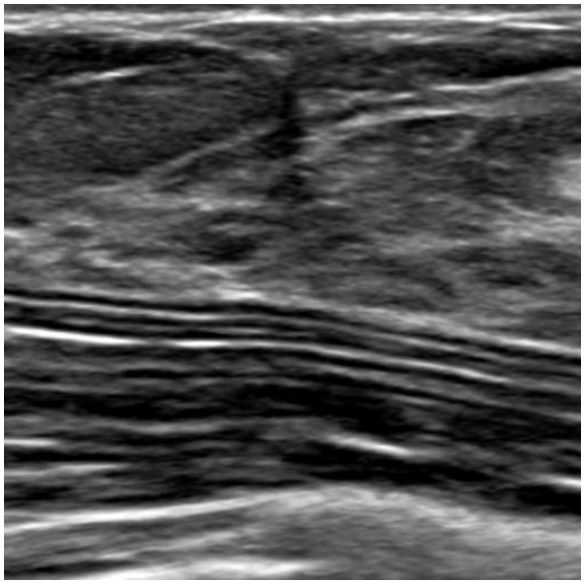


Figure 4.10. Examples of upsampled test images. The images were upsampled using (a)Bicubic interpolation (b) Modified SRGAN (c) Modified SRGAN with histogram equalization (d) Original HR image

Figure 4.7, Figure 4.8, Figure 4.9 and Figure 4.10 have examples of upsampled test images along with their corresponding MSE and NIQE. In Figure 4.7 and Figure 4.8, we compare the image upsampled using bicubic interpolation, the original SRGAN network and the modified SRGAN to its original HR image. In Figure 4.9 and Figure 4.10, we compare the image upsampled using bicubic interpolation, the modified SRGAN network and the modified SRGAN with histogram equalization to its original HR image.

4.5 DISCUSSION

In our work, we evaluated the performance of the SRGAN for upsampling ultrasound images. The SRGAN was trained on ultrasound images of the breast. Training the original SRGAN network generated checkerboard artifacts illustrated in Figure 4.6. Odena et al. have shown that for other neural networks replacing the deconvolution layer can remove this artifact [82]. We have shown that replacing the deconvolution layer with a nearest neighbor interpolation operation and a convolutional layer removes the checkerboard artifact for the SRGAN trained on ultrasound images. Modifying the SRGAN reduces the MSE of the upsampled image compared to using the original architecture of the SRGAN. This is depicted in Table 4.1 and Figure 4.3. Although the images upsampled using bicubic interpolation have a lower MSE compared to the different versions of the SRGAN, MSE is not the best indicator of image quality. As illustrated in Figure 4.7 and Figure 4.8, the example images upsampled by the modified SRGAN looked closer to the HR image compared to the image upsampled using bicubic interpolation yet the image upsampled using bi-cubic interpolation has a lower MSE. Note that the NIQE in this case is a much better indicator of image quality. As illustrated in Figure 4.5, all the test images upsampled using the modified SRGAN have a lower NIQE indicating better image quality compared to images

upsampled using bicubic interpolation. The average PSNR computed on each of the different upsampling schemes are comparable as reported in Table 4.1. The PSNR appears to improve when using the modified SRGAN over the original SRGAN as illustrated in Figure 4.4.

Applying a histogram equalization algorithm on the images upsampled using the SRGAN appears to have a mixed effect as illustrated in in Figure 4.9 and Figure 4.10. In Figure 4.9, histogram equalization appears to have increased the MSE and the NIQE while in Figure 4.10, histogram equalization appears to have reduced the MSE but the NIQE has increased. This is also evident in Figure 4.3 and Figure 4.5. If we compare Figure 4.3, Figure 4.4 and Figure 4.5 histogram equalization appears to have increased MSE, reduced PSNR and increased the NIQE compared to bicubic interpolation. It appears histogram equalization did not improve the performance of the SRGAN.

4.6 CONCLUSION

We evaluated the performance of the SRGAN for upsampling ultrasound breast images by a factor of 4. We compared the performance of the SRGAN algorithm to bicubic interpolation. Although we compared metrics like the MSE, PSNR and NIQE, none of these are perfect predictors of image quality. A better way to quantify the performance of the algorithm would be to have human observers' rate each of the upsampled images, or even better, would be to use images for a specific task and determine which technique provides the best task-based performance. We evaluated the SRGAN for upsampling images by a factor of 4 only. As part of our future work, we could evaluate the performance of the algorithm using different scaling factors and also on ultrasound images of a different anatomy to see if this might affect the algorithm.

Chapter 5. CONCLUSION

In our work, we compared different CNN based approaches for medical image analysis tasks. We considered two image classification tasks and an image enhancement task. The first task involves characterizing thyroid nodules using its B-mode ultrasound and SWE image. The second task involves identifying the presence of breast lesions using its corresponding MIP DCE-MRI image. For the classification tasks we compared the performance of different CNN based classifiers: fine tuning a pre-trained the ResNet50 architecture[42], training a one-class auto encoder and a Siamese neural network. The one-class autoencoder and the Siamese neural network were chosen because of their ability to learn from an unbalanced and much smaller data set, common attriburtes of medical data. We also applied techniques such as weighting the samples during training, adding a drop out layer, applying an early stopping criterion and regularizing the weights of the network to help with the unbalanced and smaller size of our data set. For characterization of thyroid nodules using SWE images, we obtained accuracies ranging from 80% - 87% which is comparable to reported human diagnostic accuracy [40]. The highest AUC we achieved using the B-mode image was 0.65. This was much lower than the AUC reported by similar studies [47]–[49] possibly because of compression artifacts present on the B-mode image [35]. Also, the segmented images of the thyroid lesion contained a lot of background tissue which could affect the performance of the classification algorithm. For the second task of identifying breast lesions using its MIP DCE-MRI image, we obtained the best performance (AUC = 0.73) by using the pre-trained the ResNet [42]. The one class auto encoder and the Siamese neural network gave us an AUC of 0.54 and 0.5 respectively. Other groups have reported a higher AUC of 0.76 to 0.88 [63]–[66], but these were for the task of classifying lesions as malignant or benign. Furthermore, these other approaches used segmented lesions

(images of only the lesion, with limited background). In our work, we trained the network on images of the entire left or right breast instead of training the network on segmentations of the breast lesion.

For the third task of upsampling the B-mode ultrasound images using the SRGAN, we reported the artifacts that were generated. We modified the architecture to correct these artifacts. Finally, we compared its performance to a standard bicubic interpolation algorithm using metrics such as the MSE, the PSNR and the NIQE. Initial results suggest that the SRGAN provides high-resolution images with similar resolution and noise texture as the original high-resolution images, although further work is needed to determine if this approach translates to other noise levels and ultrasound techniques.

5.1 Future Work

For each of the two medical image classification tasks, we could train the networks on segmented images of the thyroid and breast lesion respectively. We could also integrate clinical information into the prediction algorithm. Both these steps might improve the performance of the deep learning-based classification algorithms. For the deep learning based upsampling schemes, we could have human observers quantify the image quality of the different upsampling schemes. We could evaluate the performance of the algorithm using different scaling factors and on ultrasound images of a different anatomy. Currently the SRGAN has been trained and evaluated on ultrasound images of the breast only.

In summation, this work demonstrates the use of deep learning methods for common medical imaging tasks of classification and image enhancement. We applied recent developments in deep learning to real medical imaging data, containing the common challenges in this space of

limited data volumes and unbalanced class sets. Despite these challenges, we demonstrate that several architectures can be effectively trained and applied with reasonable performance.

BIBLIOGRAPHY

[1] P. Domingos, “A Few Useful Things to Know about Machine Learning.”

[2] B. J. Erickson, P. Korfiatis, Z. Akkus, and T. L. Kline, “Machine Learning for Medical Imaging,” *RadioGraphics*, vol. 37, no. 2, pp. 505–515, Mar. 2017.

[3] M. A. Mazurowski, M. Buda, A. Saha, and M. R. Bashir, “Deep learning in radiology: an overview of the concepts and a survey of the state of the art,” Feb. 2018.

[4] J. Gerrity, “Comment: Health networks - delivering the future of healthcare,” *Building Better Health Care*, 2014. [Online]. Available: https://www.buildingbetterhealthcare.co.uk/technical/article_page/Comment_Health_networks__delivering_the_future_of_healthcare/94931. [Accessed: 22-Sep-2018].

[5] D. Ribli, A. Horváth, Z. Unger, P. Pollner, and I. Csabai, “Detecting and classifying lesions in mammograms with Deep Learning,” *Sci. Rep.*, vol. 8, no. 1, p. 4165, Dec. 2018.

[6] B. Kayalibay, G. Jensen, and P. van der Smagt, “CNN-based Segmentation of Medical Imaging Data,” Jan. 2017.

[7] Y. Zhang and H. Yu, “Convolutional Neural Network based Metal Artifact Reduction in X-ray Computed Tomography.”

[8] M. Z. Alom *et al.*, “The History Began from AlexNet: A Comprehensive Survey on Deep Learning Approaches,” Mar. 2018.

[9] H. Wang and B. Raj, “On the Origin of Deep Learning,” Feb. 2017.

[10] A. L. Beam, “Deep Learning 101 - Part 1: History and Background.” [Online]. Available: https://beamandrew.github.io/deeplearning/2017/02/23/deep_learning_101_part1.html.

[Accessed: 11-Nov-2018].

[11] J. Sandhu, “A Concise History of Neural Networks – Jiaconda – Medium,” *2016-08-13*. [Online]. Available: https://medium.com/@Jaconda/a-concise-history-of-neural-networks-2070655d3fec. [Accessed: 10-Nov-2018].

[12] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning representations by back-propagating errors,” *Nature*, vol. 323, no. 6088, pp. 533–536, Oct. 1986.

[13] J. Schmidhuber, “Who Invented Backpropagation?,” 2015. [Online]. Available: http://people.idsia.ch/~juergen/who-invented-backpropagation.html. [Accessed: 11-Nov-2018].

[14] A. Karpathy, “CS231n Convolutional Neural Networks for Visual Recognition.” [Online]. Available: http://cs231n.github.io/neural-networks-1/. [Accessed: 25-Sep-2018].

[15] A. Singh, “Activation functions and it’s types-Which is better?,” 2017. [Online]. Available: https://towardsdatascience.com/activation-functions-and-its-types-which-is-better-a9a5310cc8f. [Accessed: 11-Nov-2018].

[16] P. Seeböck, “Deep Learning in Medical Image Analysis,” Vienna University of Technology, 2015.

[17] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[18] D. P. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization,” Dec. 2014.

[19] J. Brownlee, “What is the Difference Between Test and Validation Datasets?,” 2017. [Online]. Available: https://machinelearningmastery.com/difference-test-validation-datasets/. [Accessed: 11-Nov-2018].

[20] H. Greenspan, B. van Ginneken, and R. M. Summers, “Guest Editorial Deep Learning in Medical Imaging: Overview and Future Promise of an Exciting New Technique,” *IEEE Trans. Med. Imaging*, vol. 35, no. 5, pp. 1153–1159, May 2016.

[21] S.-C. B. Lo, J.-S. Lin, M. T. Freedman, and S. K. Mun, “Computer-assisted diagnosis of lung nodule detection using artificial convolution neural networ,” 1993, vol. 1898, pp. 859–869.

[22] H.-P. Chan, S.-C. B. Lo, B. Sahiner, K. L. Lam, and M. A. Helvie, “Computer-aided detection of mammographic microcalcifications: Pattern recognition with an artificial neural network,” *Med. Phys.*, vol. 22, no. 10, pp. 1555–1567, Oct. 1995.

[23] B. Sahiner *et al.*, “Classification of mass and normal breast tissue: a convolution neural network classifier with spatial domain and texture images,” *IEEE Trans. Med. Imaging*, vol. 15, no. 5, pp. 598–610, 1996.

[24] J. Schmidhuber, *Deep Learning in Neural Networks: An Overview*. 2014.

[25] C. Xiao, E. Choi, and J. Sun, “Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review,” *J. Am. Med. Informatics Assoc.*, Jun. 2018.

[26] N. Srivastava, G. Hinton, A. Krizhevsky, and R. Salakhutdinov, “Dropout: A Simple Way to Prevent Neural Networks from Overfitting,” 2014.

[27] J. Wang and L. Perez, “The Effectiveness of Data Augmentation in Image Classification using Deep Learning.”

[28] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-Image Translation with Conditional Adversarial Networks,” Nov. 2016.

[29] I. J. Goodfellow *et al.*, “Generative Adversarial Networks,” Jun. 2014.

[30] M. C. Frates *et al.*, “Management of Thyroid Nodules Detected at US: Society of Radiologists in Ultrasound Consensus Conference Statement,” *Radiology*, vol. 237, no. 3, pp. 794–800, Dec. 2005.

[31] A. Jemal *et al.*, “Cancer statistics, 2005.,” *CA. Cancer J. Clin.*, vol. 55, no. 1, pp. 10–30.

[32] F. N. Tessler *et al.*, “ACR Thyroid Imaging, Reporting and Data System (TI-RADS): White Paper of the ACR TI-RADS Committee,” *J. Am. Coll. Radiol.*, vol. 14, no. 5, pp. 587–595, May 2017.

[33] C. Xie, P. Cox, N. Taylor, and S. LaPorte, “Ultrasonography of thyroid nodules: a pictorial review.,” *Insights Imaging*, vol. 7, no. 1, pp. 77–86, Feb. 2016.

[34] B. R. Haugen *et al.*, “2015 American Thyroid Association Management Guidelines for Adult Patients with Thyroid Nodules and Differentiated Thyroid Cancer The American Thyroid Association Guidelines Task Force on Thyroid Nodules and Differentiated Thyroid Cancer,” *THYROID*, vol. 26, no. 1, 2016.

[35] M. Dighe, D. S. Hippe, and J. Thiel, “Artifacts in Shear Wave Elastography Images of Thyroid Nodules,” *Ultrasound Med. Biol.*, vol. 44, no. 6, pp. 1170–1176, Jun. 2018.

[36] J. Bercoff, M. Tanter, and M. Fink, “Supersonic shear imaging: a new technique for soft tissue elasticity mapping.,” *IEEE Trans. Ultrason. Ferroelectr. Freq. Control*, vol. 51, no. 4, pp. 396–409, Apr. 2004.

[37] F. Sebag *et al.*, “Shear Wave Elastography: A New Ultrasound Imaging Mode for the Differential Diagnosis of Benign and Malignant Thyroid Nodules,” *J. Clin. Endocrinol. Metab.*, vol. 95, no. 12, pp. 5281–5288, Dec. 2010.

[38] G. Azizi, J. Keller, M. Lewis, D. Puett, K. Rivenbark, and C. Malchoff, “Performance of Elastography for the Evaluation of Thyroid Nodules: A Prospective Study,” *Thyroid*, vol. 23, no. 6, pp. 734–740, Jun. 2013.

[39] C. Asteria *et al.*, “US-Elastography in the Differential Diagnosis of Benign and Malignant Thyroid Nodules,” *Thyroid*, vol. 18, no. 5, pp. 523–531, May 2008.

[40] V. Veer and S. Puttagunta, “The role of elastography in evaluating thyroid nodules: a literature review and meta-analysis,” *Eur. Arch. Oto-Rhino-Laryngology*, vol. 272, no. 8, pp. 1845–1855, Aug. 2015.

[41] C. Pereira, M. K. Dighe, and A. M. Alessio, “Comparison of machine learned approaches

for thyroid nodule characterization from shear wave elastography images," in *Medical Imaging 2018: Computer-Aided Diagnosis*, 2018, vol. 10575, p. 68.

[42] K. He, X. Zhang, S. Ren, and J. Sun, *Identity Mappings in Deep Residual Networks*. .

[43] D. Koundal, "Computer-Aided Diagnosis of Thyroid Nodule: A Review," *Int. J. Comput. Sci. Eng. Surv.*, vol. 3, no. 4, pp. 67–83, 2012.

[44] M. Savelonas, D. Maroulis, and M. Sangriotis, "A computer-aided system for malignancy risk assessment of nodules in thyroid US images based on boundary features," *Comput. Methods Programs Biomed.*, vol. 96, no. 1, pp. 25–32, Oct. 2009.

[45] Y. Chang *et al.*, "Computer-aided diagnosis for classifying benign versus malignant thyroid nodules based on ultrasound images: A comparison with radiologist-based assessments," *Med Phys*, vol. 43, no. 1, p. 554, Jan. 2016.

[46] Q. Yu, T. Jiang, A. Zhou, L. Zhang, C. Zhang, and P. Xu, "Computer-aided diagnosis of malignant or benign thyroid nodes based on ultrasound images," *Eur. Arch. Oto-Rhino-Laryngology*, vol. 274, no. 7, pp. 2891–2897, Jul. 2017.

[47] J. Ma, F. Wu, J. Zhu, D. Xu, and D. Kong, "A pre-trained convolutional neural network based method for thyroid nodule diagnosis," *Ultrasonics*, vol. 73, pp. 221–230, Jan. 2017.

[48] J. Chi, E. Walia, P. Babyn, J. Wang, G. Groot, and M. Eramian, "Thyroid Nodule Classification in Ultrasound Images by Fine-Tuning Deep Convolutional Neural Network," *J. Digit. Imaging*, vol. 30, no. 4, pp. 477–486, Aug. 2017.

[49] T. Liu, S. Xie, Y. Zhang, J. Yu, L. Niu, and W. Sun, "Feature selection and thyroid nodule classification using transfer learning," in *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, 2017, pp. 1096–1099.

[50] "MATLAB 2018b." The MathWorks, Inc., Natick, Massachusetts, United States.

[51] Q. Wei, B. Shi, J. Y. Lo, L. Carin, Y. Ren, and R. Hou, "Anomaly detection for medical images based on a one-class classification," in *Medical Imaging 2018: Computer-Aided*

Diagnosis, 2018, p. 57.

[52] P. Saduikis, “Autoencoder,” *GitHub Repos.*, 2018.

[53] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks.” pp. 1097–1105, 2012.

[54] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” Sep. 2014.

[55] G. Koch, R. Zemel, and R. Salakhutdinov, “Siamese Neural Networks for One-shot Image Recognition.”

[56] F. Chollet, “Keras,” *GitHub Repos.*, 2015.

[57] N. Tajbakhsh *et al.*, “Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning?,” *IEEE Trans. Med. Imaging*, vol. 35, no. 5, pp. 1299–1312, May 2016.

[58] N. H. G. M. Peters, I. H. M. Borel Rinkes, N. P. A. Zuithoff, W. P. T. M. Mali, K. G. M. Moons, and P. H. M. Peeters, “Meta-Analysis of MR Imaging in the Diagnosis of Breast Lesions,” *Radiology*, vol. 246, no. 1, pp. 116–124, Jan. 2008.

[59] L. Li *et al.*, “Parameters of dynamic contrast-enhanced MRI as imaging markers for angiogenesis and proliferation in human breast cancer.,” *Med. Sci. Monit.*, vol. 21, pp. 376–82, Feb. 2015.

[60] American College of Radiology., “ACR practice parameter for the performance of contrast-enhanced magnetic resonance imaging (MRI) of the breast.”

[61] A. G. Sorace *et al.*, “Distinguishing benign and malignant breast tumors: preliminary comparison of kinetic modeling approaches using multi-institutional dynamic contrast-enhanced MRI data from the International Breast MR Consortium 6883 trial,” *J. Med. Imaging*, vol. 5, no. 01, p. 1, Jan. 2018.

[62] D. M. Ikeda *et al.*, “Development, standardization, and testing of a lexicon for reporting contrast-enhanced breast magnetic resonance imaging studies,” *J. Magn. Reson. Imaging*, vol. 13, no. 6, pp. 889–895, Jun. 2001.

[63] N. Antropova, B. Huynh, and M. Giger, “SU-D-207B-06: Predicting Breast Cancer Malignancy On DCE-MRI Data Using Pre-Trained Convolutional Neural Networks,” *Med. Phys.*, vol. 43, no. 6Part4, pp. 3349–3350, Jun. 2016.

[64] S. Marrone, G. Piantadosi, R. Fusco, A. Petrillo, M. Sansone, and C. Sansone, “An Investigation of Deep Learning for Lesions Malignancy Classification in Breast DCE-MRI,” Springer, Cham, 2017, pp. 479–489.

[65] N. Antropova, B. Q. Huynh, and M. L. Giger, “A deep feature fusion methodology for breast cancer diagnosis demonstrated on three imaging modality datasets,” *Med. Phys.*, vol. 44, no. 10, pp. 5162–5171, Oct. 2017.

[66] N. Antropova, H. Abe, and M. L. Giger, “Use of clinical MRI maximum intensity projections for improved breast lesion classification with deep convolutional neural networks,” *J. Med. Imaging*, vol. 5, no. 01, p. 1, Feb. 2018.

[67] S. C. Agner *et al.*, “Textural Kinetics: A Novel Dynamic Contrast-Enhanced (DCE)-MRI Feature for Breast Lesion Classification,” *J. Digit. Imaging*, vol. 24, no. 3, pp. 446–463, Jun. 2011.

[68] S. Pereira, A. Pinto, V. Alves, and C. A. Silva, “Brain Tumor Segmentation Using Convolutional Neural Networks in MRI Images,” *IEEE Trans. Med. Imaging*, vol. 35, no. 5, pp. 1240–1251, May 2016.

[69] B. Di Muzio, “Maximum Intensity Projection (MIP).” [Online]. Available: https://radiopaedia.org/articles/maximum-intensity-projection-mip. [Accessed: 08-Nov-2018].

[70] M. Ploquin, A. Basarab, and D. Kouamé, “Resolution enhancement in medical ultrasound imaging.,” *J. Med. imaging (Bellingham, Wash.)*, vol. 2, no. 1, p. 017001, Jan. 2015.

[71] H. Hasegawa, “Improvement of range spatial resolution of medical ultrasound imaging by element-domain signal processing,” *Jpn. J. Appl. Phys.*, vol. 56, no. 7S1, p. 07JF02, Jul. 2017.

[72] A. Ng and J. Swanevelder, “Resolution in ultrasound imaging,” *Contin. Educ. Anaesth. Crit. Care Pain*, vol. 11, no. 5, pp. 186–192, Oct. 2011.

[73] R. G. Keys, *Cubic Convolution Interpolation for Digital Image Processing*, no. 6. 1981, p. 1153.

[74] C. E. Duchon, “Lanczos Filtering in One and Two Dimensions,” *J. Appl. Meteorol.*, vol. 18, no. 8, pp. 1016–1022, Aug. 1979.

[75] W. Yang, X. Zhang, Y. Tian, W. Wang, and J.-H. Xue, *Deep Learning for Single Image Super-Resolution: A Brief Review*. .

[76] C. Dong, C. C. Loy, K. He, and X. Tang, “Image Super-Resolution Using Deep Convolutional Networks,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307.

[77] C. Dong, C. C. Loy, and X. Tang, *Accelerating the Super-Resolution Convolutional Neural Network*. .

[78] J. Kim, J. K. Lee, and K. M. Lee, “Accurate Image Super-Resolution Using Very Deep Convolutional Networks,” Nov. 2015.

[79] C. Ledig *et al.*, “Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network,” Sep. 2016.

[80] Y. Dong, Hao and Supratak, Akara and Mai, Luo and Liu, Fangde and Oehmichen, Axel and Yu, Simiao and Guo, “TensorLayer: A Versatile Library for Efficient Deep Learning Development,” *ACM Multimed.*, 2017.

[81] A. Mittal, R. Soundararajan, and A. C. Bovik, “Making a ‘Completely Blind’ Image Quality Analyzer,” *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, Mar. 2013.

[82] A. Odena, V. Dumoulin, and C. Olah, “Deconvolution and Checkerboard Artifacts,” *Distill*, 2016.