Explainable AI and susceptibility to adversarial attacks in classification and segmentation of breast ultrasound images

Hamza Rasaee

A Thesis

in

The Department

of

ELECTRICAL AND COMPUTER SCIENCE ENGINEERING

Presented in Partial Fulfillment of the Requirements

for the Degree of

Master of Applied Science (Electrical And Computer Science Engineering) at

Concordia University

Montréal, Québec, Canada

January 2022

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: **Hamza Rasaee**

Entitled: **Explainable AI and susceptibility to adversarial attacks in classification and segmentation of breast ultrasound images**

and submitted in partial fulfillment of the requirements for the degree of

Master of Applied Science (Electrical And Computer Science Engineering)

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the Final Examining Committee:

Chair
Dr. O. Ait Mohamed

External Examiner
Dr. J. Yan (CIISE)

Examiner
Dr. O. Ait Mohamed

Co-Supervisor
Dr. H. Rivaz

Co-supervisor
Dr. F. Nasiri (BCEE)

Approved by

Y. R. Shayan, Chair
Department of ELECTRICAL AND COMPUTER SCIENCE ENGINEERING

2022

Mourad Debbabi, Dean
Faculty of Engineering and Computer Science

Abstract

Explainable AI and susceptibility to adversarial attacks in classification and segmentation of breast ultrasound images

Hamza Rasaee

Ultrasound is a non-invasive imaging modality that can be conveniently used to classify suspicious breast nodules and potentially detect the onset of breast cancer. Recently, Convolutional Neural Networks (CNN) techniques have shown promising results in classifying ultrasound images of the breast into benign or malignant. However, CNN inference acts as a black-box model, and as such, its decision-making is not interpretable. Therefore, increasing effort has been dedicated to explaining this process, most notably through Gradient-weighted Class Activation Mapping (Grad-CAM) and other techniques that provide visual explanations into inner workings of CNNs. In addition to interpretation, these methods provide clinically important information, such as identifying the location for biopsy or treatment. In this work, we analyze how adversarial assaults that are practically undetectable may be devised to alter these importance maps dramatically. Furthermore, we will show that this change in the importance maps can come with or without altering the classification result, rendering them even harder to detect. As such, care must be taken when using these importance maps to shed light on the inner workings of deep learning. Finally, we utilize Multi-Task Learning (MTL) and propose a new network based on deep residual networks to improve the classification accuracies. Our sensitivity and specificity values are comparable to the state of the art results.

Acknowledgments

First and foremost, I want to express my gratitude to my outstanding supervisors, Dr. Hassan Rivaz and Dr. Fuzhan Nasiri, for their guidance and support during my studies. I am grateful for their new ideas, proper leadership, and meticulous review, which gave me with an excellent opportunity throughout my studies. During my stay at Concordia University, I could not have wished for greater intellectuals who have continued to mentor me and never stopped caring about my achievement.

My heartfelt thanks to my family, who have always had my support despite the fact that I am thousands of miles away. Without their generous and sincere cooperation, none of this would have been possible. In addition, I appreciate my Concordia University colleagues and friends, as well as the IMPACT lab, for their thoughtful remarks and stimulating discussions.

We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC).

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The fundamentals of ultrasound imaging and related deep learning models are covered in this chapter. At the start, a brief history of ultrasound usage is presented. The physics of Ultrasound is then discussed, as well as some of its applications. Then we talk about deep learning methods, classification, segmentation, and multi-task learning in ultrasound imaging. Next, we describe a way to explain deep learning network. Finally, adversarial attack perturbations get into discussion.

1.1 Basic Physics of Ultrasound

At the beginning of the 19^{th} century, ultrasound waves were employed for navigation in submarines. It was started in the medical sector for diagnosis and therapy approximately 50 years later. Ultrasound machines have become a global imaging method with many uses with the technical advances in electronics and computing Kaproth-Joslin, Nicola, and Dogra (2015).

The Ultrasound (US) waves are sound waves that surpass the maximum limit of hearing sounds for human ears at a frequency higher than $20kHz$ Hall (2003). The US is classified as mechanical waves, i.e., by compression and expansion with longitudinal motion, moving along the material without any motions on either side Jensen (1991).

In the normal frequency range of 1 to $20MHz$, ultrasonography devices use ultrasonic waves to obtain information from the human organs. A device called transducer or probe sends the waves into the human body's area of interest. The transducer is a multi-piezoelectric crystal electrical device

Figure 1.1: An imaging ultrasound machine from Wikimedia Commons (2021)

Aldrich (2007); Narouze (2018).

The numerous designs of piezoelectric crystals and their diverse functions have resulted in a diversity of US transductors. Fig. 1.1 depicts a US machine as an instance of the diverse range of US machine models.

The transducer transforms the electrical charge into the waves of Ultrasound. This phenomenon was initially introduced in 1880 and is called a piezoelectric effect. The US waves penetrate the human body and traverse the course of transmission through diverse tissues with unique acoustic features.

Only a tiny percentage of US waves bounce back to the transducer at layer borders. At the same time, the rest continue to penetrate deeper, evaporate as heat, or disperse in various directions as depicted in Fig. 1.2. The transducer receives the reflected echoes and converts them to electrical pulses for further processing. Radio Frequency (RF) data is the technical term for the received pulses. Probes come in a variety of shapes and sizes, depending on the application and specifications; Fig. 1.3 shows four examples.

Reflected US waves can be classified in terms of frequency (measured in Hertz), wavelength (measured in Millimeter), and amplitude (measured in Decibel), just like every other type of sound wave. Each of these attributes offers useful information about the features of scanned tissue and

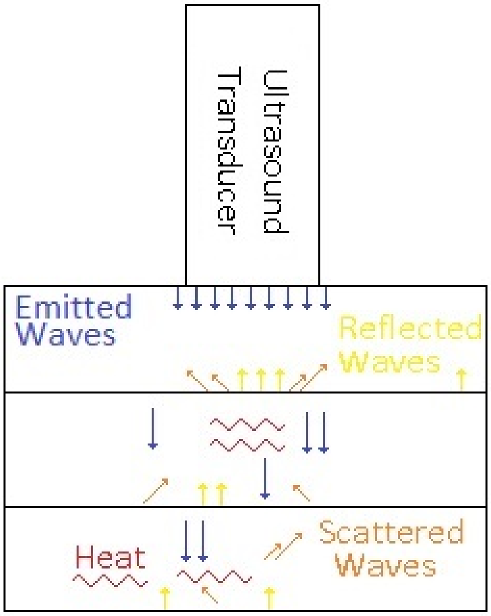


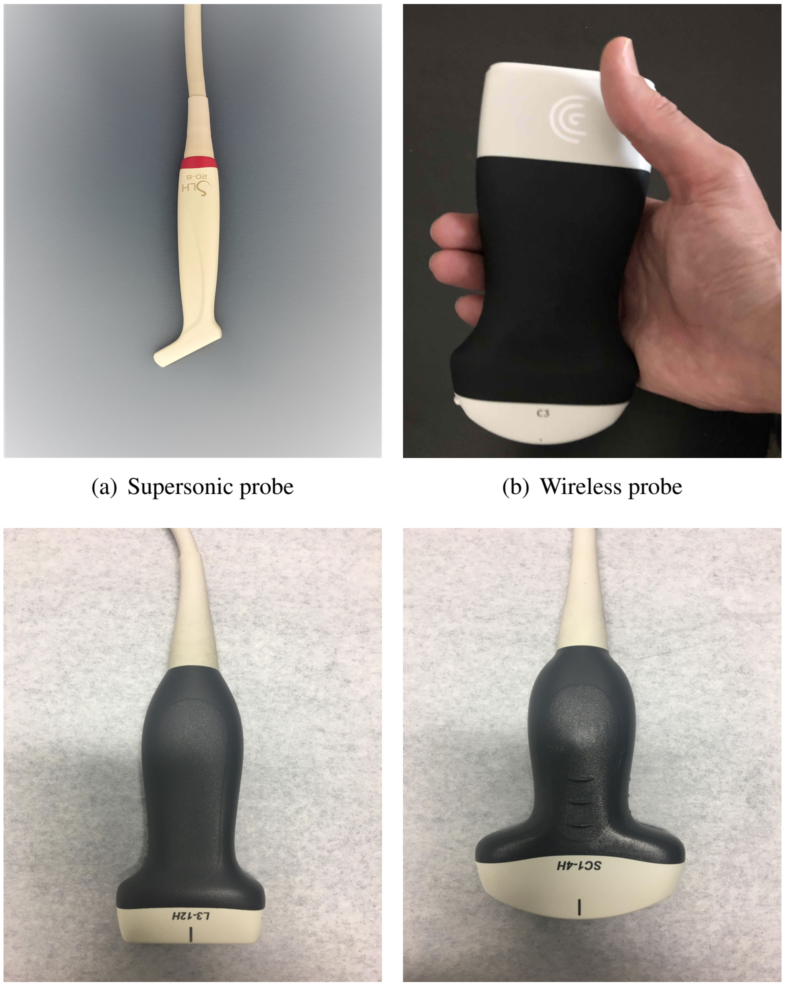
Figure 1.2: Ultrasound wave interactions with tissue

may be utilized for a variety of purposes.

The basic interpretation of US waves is based on the wave's amplitude at various time delays. The B-mode image, a gray-scale image presented on the medical US equipment interface, is essentially the envelope detection of the reflected US wave. Organs with varying densities are depicted in the B-mode image with varied brightness due to variable acoustical impedances. Acoustic impedance is calculated by multiplying the density parameter by the wave speed in the tissue Heiss et al. (1991); Shankar et al. (2001). Fig. 1.4 shows a the B-mode image of a phantoms.

Human organs have varying acoustic impedances, which results in proportional reflections and, as a result, varying brightness in the B-mode image. The lungs, which contain air, have the lowest acoustic impedance, whereas dense organs, such as bones, have the highest. A radiologist or practitioner must have a firm grasp of human anatomy and significant experience with US images to appropriately use the reproduced B-mode images to diagnose any abnormalities in the body of the patient.

Furthermore, even in a homogeneous tissue, there are usually some particles that cause partial scatterings, which causes the scattering of any sub-resolution structures to increase Burckhardt



(a) Supersonic probe

(b) Wireless probe

(c) Alpinion linear probe

(d) Alpinion convex probe

Figure 1.3: Different sample of ultrasound transducers which are available at PERFORM center

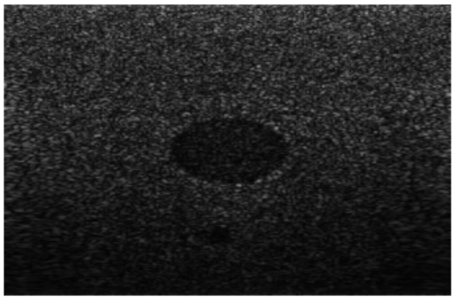


Figure 1.4: B-mode image of phantom

(1978).

In B-Mode images, these fragmentary scatterings look as speckles, potentially complicating interpretation of US images. There have been several research and attempts to reduce these scattering effects and improve the clarity of B-mode images. These components, on the other hand, define the material's microstructure.

We can learn more about the tissue characteristics by analyzing the signals reflected from these speckles. For example, we discriminate between distinct body organs and compare them to normal and healthy organs. Many studies use reflected signals to determine the location of specific organs. Segmentation is a well-known popular subject of research, which we shall return to in Section 2.1.2.

More importantly, these reflected waves might be utilized to detect and identify various anomalies in the human body, such as tumors and cysts, as well as determine whether diseased tissue is malignant or benign. This latter use is an excellent example of ultrasonic classification which we explain in Section 2.1.1.

1.2 Deep Learning

Different methodologies have been applied for analyzing US images. Dynamic programming is a computer programming technique that uses mathematical optimization Vajihi, Rosado-Mendez, Hall, and Rivaz (2018) to divide a complex issue into smaller sub-problems by using recursive functions. For example, in this paper Vajihi et al. (2018), they applied dynamic programming to estimate backscatter and attenuation on the US images. Recently with advancements in computer hardware like Graphics Processing Unit (GPU) and multiprocessors, another method which is known as deep learning LeCun, Bengio, and Hinton (2015) calculates very complex mathematical algorithms fast. Deep learning applications can be divided into single-task learning and multitask learning. In single-task learning, the model is focused on achieving a single objective (for example, one model for classification and another for segmentation) whereas, in multitask learning, one model predicts many goals (one model to classify and segment US images).

Different methods has been applied for analyzing US images. Dynamic programming is a computer programming approach that recursively divides a complex issue into simpler sub-problems using mathematical optimization Bellman (1966). In paper Vajihi et al. (2018), they applied dynamic programming to estimate backscatter and attenuation on the US images. Recently with advancements in computer hardware like Graphics Processing Unit (GPU) and multiprocessors, another method which is known as deep learning LeCun et al. (2015) that calculates very complex mathematical algorithms fast. In Chapter 2 and Chapter 3 we get more details in these regards.

1.2.1 Single Task Learning Classification In Ultrasound Imaging

Deep learning has become widely employed in medical imaging applications due to its success in computer vision. Ultrasound imaging has been one of the imaging applications which widely benefited from deep learning advances. For instance, the application of Ultrasound for differentiating between malignant and benign tumors in breast imaging Kornecki (2011) has significantly been developed by deep learning methods.

Even while other machine learning methods for classification might do this job without utilizing deep learning Cruz and Wishart (2006); Krishnan, Banerjee, Chakraborty, Chakraborty, and Ray (2010); Vishrutha and Ravishankar (2015), it has been demonstrated that deep learning frameworks provide superior results J.-Z. Cheng et al. (2016); Han et al. (2017a); Jalalian et al. (2013).

Deep learning is presently being utilized to create predictions by computer-aided diagnosis (CADx) systems, which are used to offer an objective report to assist radiologists with the interpretation and diagnosis of the medical image J.-Z. Cheng et al. (2010); Drukker, Sennett, and Giger (2008); Giger, Karssemeijer, and Schnabel (2013); Van Ginneken, Schaefer-Prokop, and Prokop (2011).

Differentiating cancerous from non-cancerous tumors is regarded as one of the most significant uses of CADx systems due to the high risk involved with categorizing a malignant tumor as benign J.-Z. Cheng et al. (2010); T. Sun, Zhang, Wang, Li, and Guo (2013); J. Wang et al. (2016); Way et al. (2006).

Convolution layers have opened the way for extracting the most valuable features using recent state-of-the-art deep learning approaches. Deep learning approaches' promising results in computer vision classification tasks have piqued researchers' interest in applying them to medical image classification Antropova, Huynh, and Giger (2017); Becker et al. (2018); Esteva et al. (2017); Han et al.

(2017b); Ting, Tan, and Sim (2019).

Deep learning techniques for breast lesion classification in mammography and US images were proposed by Esteva et al. (2017) and Han et al. (2017b). B-mode images are commonly used in breast classification utilizing deep learning techniques in the US. On the other hand, B-mode images contain far less information than the raw RF data from which they were created. To solve this problem, researchers at Jarosik, Klimonda, Lewandowski, and Byra (2020) have looked into the potential of utilizing RF data to classify benign and malignant tumors. However, they only retrieved 2D patches of RF data from the mass breast region in their study.

1.2.2 Single Task Learning Segmentation In Ultrasound Imaging

The quality of data has a significant impact on ultrasound image segmentation. Attenuation, speckle, shadows, and signal dropout are common artifacts that hamper the segmentation job; owing to the orientation dependency of acquisition, this might result in missing borders. The fact that the contrast between regions of interest is frequently minimal adds to the complexities. However, recent advancements in transducer design, spatial/temporal resolution, digital systems, portability, and other areas have substantially enhanced the quality of information obtained from an ultrasound instrument Noble and Boukerroui (2006).

Deep learning may also be used to automate ultrasound image segmentation Behboodi and Rivaz (2019) saving time and effort. Moreover, generative adversarial networks Goudarzi, Asif, and Rivaz (2020) were used to improve the resolution of the ultrasound image without affecting the frame rate by using a multi-focus image I. Goodfellow et al. (2014).

VGGNet, ResNet, and DenseNet were used to classify benign and malignant lesions in B-mode US data by Moon et al. (2020). They employed B-mode, segmented tumor, and segmented map as three channels of inputs to their networks.

Furthermore, the UNet design Ronneberger, Fischer, and Brox (2015) which is based on the fully convolutional network Long, Shelhamer, and Darrell (2015) is the most well-known architecture for biomedical image segmentation, utilizing many convolutional, max-pooling, and upsampling layers. Also, U-net has recently been recommended for use on simulated US images by Nair, Tran, Reiter, and Bell (2018).

1.2.3 Multi-Task Learning

In a variety of computer vision applications, such as image classification and semantic segmentation, convolutional neural networks (CNNs) have shown significant improvements. These networks, on the other hand, are usually designed to do a single purpose. A network that can execute several jobs concurrently is considerably more desired than constructing a collection of separate networks, one for each task, for more comprehensive vision systems in real-world applications. It is efficient not just in terms of memory and inference time but also in terms of data because linked jobs may have visually relevant characteristics in typical Liu, Johns, and Davison (2019).

For instance, in Y. Sun, Wang, and Tang (2013), the estimation of five facial landmarks consisting of three phases to return to the position of landmarks from coarse to fine is carried by multiple deep convolutionary neural networks. Moreover, Z. Zhang, Luo, Loy, and Tang (2014) has incorporated deep multi-task networks that improve further enhance performance to detect landmarks.

In a network for 3D automated breast ultrasound, Y. Zhou et al. (2021) presented a multi-task learning technique for the tumors to train segmentation and classification jointly. An encoder-decoder network and a lightweight multi-scale network are used in the proposed segmentation and classification technique. For classification and segmentation of tumor, they used VNet as the backbone network Milletari et al. (2016). Features derived from the encoding route are used in both segmentation and classification tasks.

Another research in 3D ultrasound imaging focuses on fetal brain alignment utilizing multi-task learning for fully automated alignment Namburete, Xie, Yaqub, Zisserman, and Noble (2018). They propose an automated technique for fetal brain alignment that relies solely on sonographic image signatures at any gestational stage. To normalize image volumes to a reference space, they employ estimated affine transformations.

1.3 Explainable AI

With the remarkable developments in deep learning, it is critical to deciphering what a model is saying. The models must be transparent to create confidence in intelligent systems and progress toward their meaningful integration. The goal of transparency is to explain why the model predicts specific outcomes. Selvaraju et al. (2017) presented a method for providing "visual explanations" for decisions made by a broad class of Convolutional Neural Network (CNN)-based models to increase their transparency. Gradient-weighted Class Activation Mapping (Grad-CAM) is a technique that employs the gradients of any target idea to produce a coarse localization map that emphasizes the critical places in the image for predicting the concept. We discuss more details in Chapter 4. T. He et al. (2019) present a new medical MLP (MediMLP) that uses Grad-CAM Selvaraju et al. (2017) (a variation of CAM) to conduct postoperative complication prediction (PCP) tasks and extract important factors for lung cancer PCP.

1.4 Susceptibility To Adversarial Attacks

Despite advances in deep learning, human and machine perception systems are still vastly different. As demonstrated by Szegedy et al. (2013), small but well-controlled visual disturbances can lead to erroneous classification in artificial systems with great confidence. On the other hand, these disruptions are usually imperceptible by humans and do not raise any doubts regarding the correct classification. For instance, adversarial examples are differentiated by requiring only small perturbations that are almost imperceptible to a human observer Metzen, Genewein, Fischer, and Bischoff (2017). In Moosavi-Dezfooli, Fawzi, and Frossard (2016), they proposed an algorithm, DeepFool, to calculate adversarial attacks that mislead modern classifiers like LeNet (MNIST), FC500-150-10 (MNIST), NIN (CIFAR-10), and LeNet (CIFAR-10). It is based on an iterative classifier linearization that results in little disruption, adequate to modify classification labels.

1.5 Problem Statement

It is crucial to trust a deep learning model, especially in the medical area. Even though Grad-CAM makes it feasible to explain a model (see Chapter 4), it still has to be evaluated with various data quality. Different noise sources can contaminate ultrasonic images, and the look of these images can be dramatically altered by adopting a different frequency or beamforming method. These modifications have the potential to skew the classification findings or the Grad-CAM. Furthermore, the predictions of the deep learning networks are sensitive to adversarial attacks I. J. Goodfellow, Shlens, and Szegedy (2014); Kurakin, Goodfellow, Bengio, and others (2016); Moosavi-Dezfooli et al. (2016).

1.6 Research Objective

The primary goal of the current thesis study is to investigate breast US images in both classification and segmentation tasks. Thus, as we explain in Chapter 4, GRAD-CAM helps us better understand and explain network activity. To that purpose, we use GRAD-CAM heat maps in this thesis to show the performance of our proposed classification model and offer human-readable reasons for our decisions.

Based on Finlayson, Chung, Kohane, and Beam (2018), we must also examine how adversarial assaults may create new potential for fraud and injury as we increase the use of AI in the medical field and remove doctors from the decision-making loop. These types of attacks could happen by cyber attack. Therefore in another objective of this research, we use adversarial assaults on input images in Chapter 5 to show that adversarial examples in breast US images can be used to manipulate CNN-based networks. As a result, we recommend that researchers, particularly in medical US imaging, be aware of existing CNN-based network weaknesses and highlight research organizations' concerns in future research of medical, educational environments.

1.7 Thesis Outline

The thesis is organized in the following order: In Chapter 2, we suggest deep learning techniques for ultrasound image segmentation and classification based on the residual network. Then in Chapter 3 we continue the work on a multi-task learning method to gain the classification and segmentation prediction results. After, in Chapter 4, we introduce a method to explain our deep neural networks which is published in Rasaee and Rivaz (2021). Chapter 5 is where we stress our designed network by applying adversarial attracts. Finally, we wrap up the thesis with conclusions and future work in Chapter 6.

Chapter 2

Single Task Learning

2.1 Introduction

This chapter begins with a brief introduction to deep learning classification 2.1.1 and segmentation 2.1.2 followed by database explanation in Section 2.2.1. Then the networks based on the ResNet-50 are explained in Sections 2.2.2, 2.2.3 along with their architecture details. In Section 2.2.4, we talk about hyperparameters with some of their technical details and how to tune these parameters. Section 2.3 explains the result of the different tasks with corresponding figures. Finally, we conclude the discussed topics in this chapter in Section 2.4.

2.1.1 Classification

There are numerous well-known classification models for single-task learning (STL), such as AlexNet, VGG16, and ResNet. VGG16 is a proper sample of a single task convolutional neural network model, which was proposed by Simonyan and Zisserman (2014) from the University of Oxford. The VGG16 model achieves 92.7% top-five test accuracy in ImageNet, a dataset that includes more than 14 million images belonging to 1000 classes J. Deng et al. (2009). This model was selected as the best performing model in ILSVRC-2014 competition. It exceeds AlexNet Krizhevsky, Sutskever, and Hinton (2012) by using a sequence of 33 kernel-sized filters instead of having a extensive kernel-sized filters 11 for the first layer and 5 for second convolutional layers.

In Lazo, Moccia, Frontoni, and De Momi (2020), they applied VGG16 and Inception-V3 models

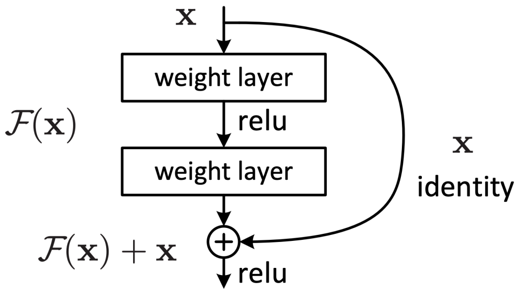


Figure 2.1: Residual block from K. He et al. (2016)

on the same dataset as we use in this thesis. Then by tuning the hyperparameter, the AUC score improved from 79.1% to 93.4% for VGG16, and from 62.3% to 78.3% for Inception-V3 model.

Another STL model is ResNet-50 K. He et al. (2016) that will be discussed in depth in this thesis. ResNet-50 is a deep residual network with a layer count of 50. The input image size is 224 by 224 pixels with three channels. ResNet is the most often used subclass of convolutional neural networks for image classification. ResNet's major innovation is the skip connection. On the other hand, deep networks are widely recognized for having vanishing gradients, which means that as the model backpropagates, the gradient grows less and smaller. Learning might be difficult due to small gradients. In Fig. 2.1, the skip connection is called "identity". It enables the network to learn the identity function, allowing the input to bypass the other weight layers and flow through the block.

2.1.2 Segmentation

In image segmentation, two primary models are well-known: U-Net and V-Net. Ronneberger et al. (2015) created the U-Net for biomedical image segmentation. There are two steps in the U-shape architecture. The encoder (contraction path) is the first path, and it is used to record the image's context. The encoder is simply a convolutional and maximum pooling layer stack. The second approach is the decoder (symmetric expanding), which uses transposed convolutions to provide accurate localization. As a result, it is an end-to-end fully convolutional network, meaning that it only has convolutional layers and no Dense layers that allow it to accept images of any size.

Milletari et al. (2016) proposed using 3D convolutions instead of processing the input 3D volumes slice-by-slice. V-Net is similar to U-Net; however, there are a few differences. It consists

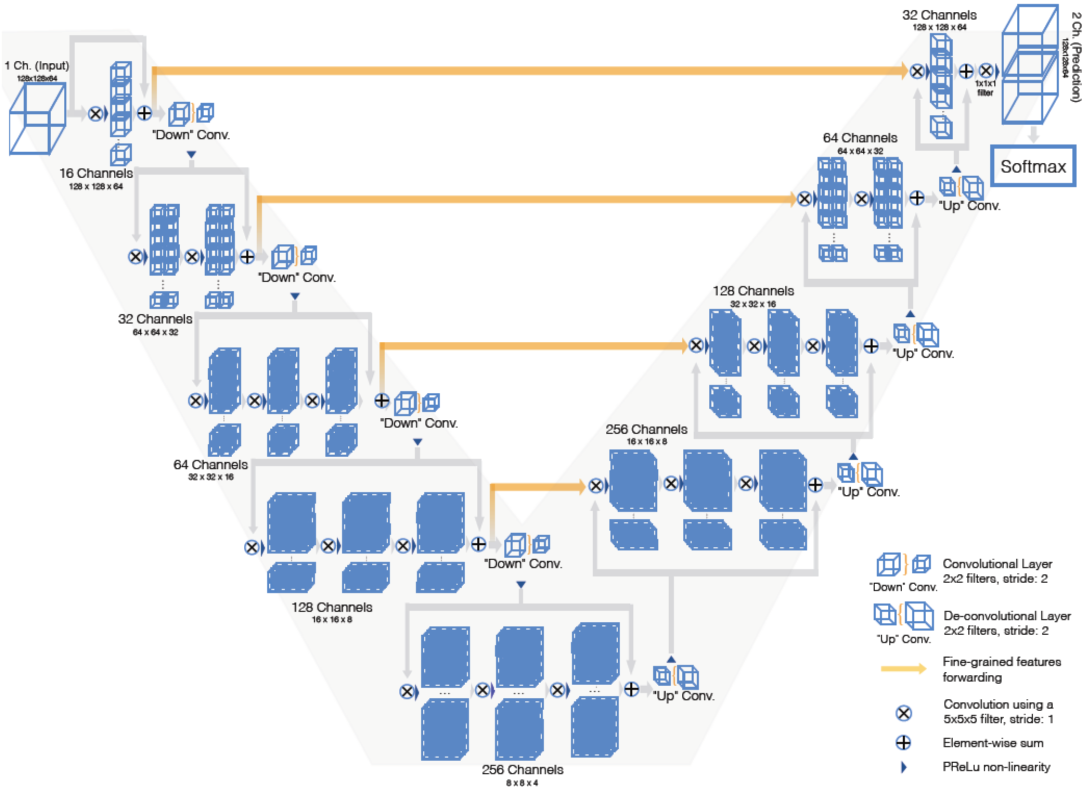


Figure 2.2: V-Net architecture from Milletari et al. (2016)

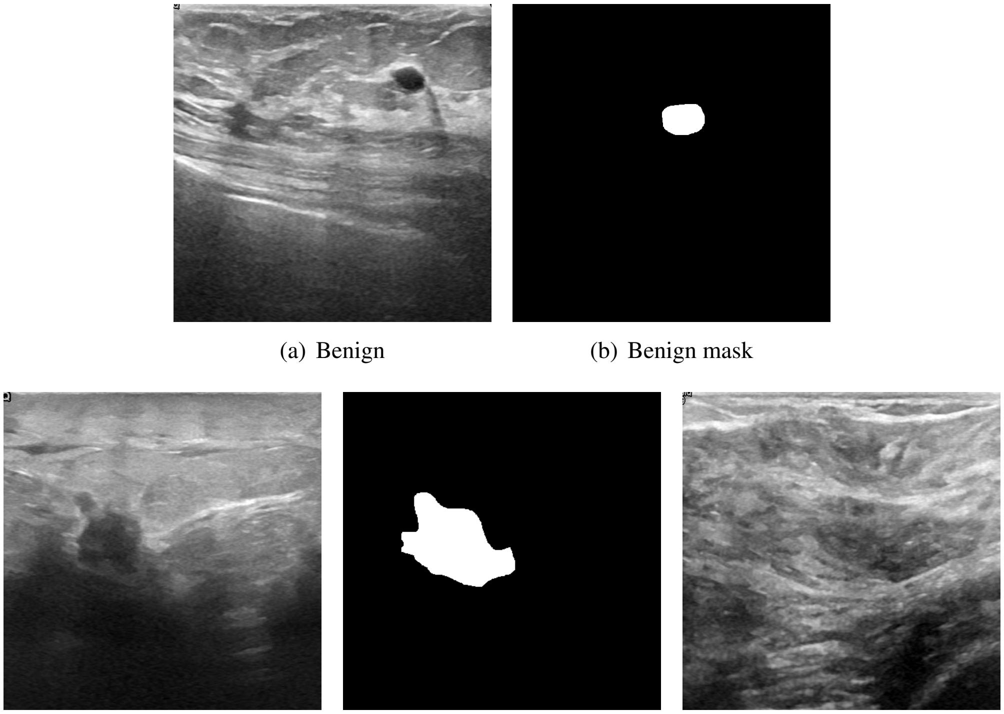
of two parts as depicted in 2.2: left and right. The network's left side is a compression path (encoder), while the right side decompresses (decoder) the signal till it reaches its original size. The network collects features and increases the spatial support of lower resolution feature maps in the right portion to gather and combine the essential information to produce a two-channel volumetric segmentation.

2.2 Method

Section 2.1 explains some famous models regarding image classification and segmentation. On the one side, to have a proper model to classify the ultrasound images, we use ResNet-50 as one of the best models to classify a wide range of images. On the other side, U-Net, with embedding Encoder (downsample) and Decoder (upsample) structure in its network, works pretty well for segmentation. Therefore, we combined the Decoder part of the U-Net to ResNet-50 model to build the segmentation network.

2.2.1 Dataset

In this thesis, we use two public datasets: The first database is retrieved from a public database Al-Dhabyani, Gomaa, Khaled, and Fahmy (2020) comprised of breast ultrasound images in PNG format, which was first gathered in 2018. The data is recorded from 600 female patients ranging in age from 25 to 75 years old. There are 780 images in the collection, with an average size of 500 × 500 pixels. These images are categorized into three groups: 437 benign, 210 malignant, and 133 normal with the related mask. So in all the tests, whenever needs the mask, we use this dataset.



(a) Benign (b) Benign mask

(c) Malignant (d) Malignant mask (e) Normal

Figure 2.3: Benign, malignant, and normal images from the first database

The second database contains 250 BMP images of breast cancer, which are split into 100 benign and 150 malignant images. The images are 72 × 72 pixels in size, with width ranging from 57 to 61 pixels and heights ranging from 75 to 199 pixels Rodrigues (2017). Due to the lack of masks in this dataset, we only use this dataset for the models designed for classification without a mask.

The dataset has two issues for training in deep learning: first, the image size for feeding the

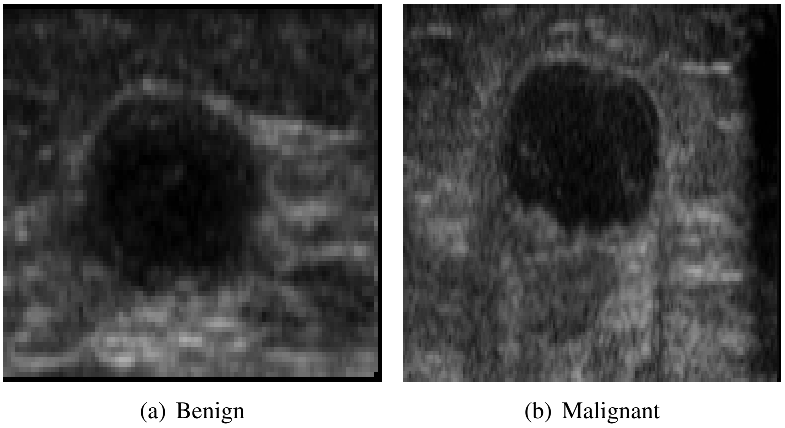

(a) Benign (b) Malignant

Figure 2.4: Benign and malignant images from the second database

model is varied. Therefore we scale the photos to 224 × 224 pixels. The second issue is the limited amount of training data. As a result, we employ a data generator for augmentation approach that includes horizontal flip, 5-degree rotation range, and a 10% height shift range. One sample of data augmentation is shown in 2.5. Then three images are produced for each image by using a data generator (totally four images: one origin image and three generated images). In the end, 70% of the dataset is used for training, 10% for validation, and 20% for testing. We take care of data leakage by setting 70% to train, 10% to validate, and 20% for testing to ensure no images from the training set would not exist in the test set.

2.2.2 Modifying ResNet-50 For Classification

Herein, we utilize a CNN model based on the ResNet-50 for the classification of breast ultrasound images. The ResNet-50 model is modified to classify input images as benign and malignant. So, the fully-connected layer with Dense-2 in ResNet-50 is replaced with fully-connected-1000 so that by using soft-max, the model could make the final binary decision. This network is illustrated in Fig. 2.6.

We apply this network 2.6 on two different types of input images. In the first experiment, we use the combination of both datasets 2.3 and 2.4 that included totally 897 images. The model accepts input images with the size of 224 × 224 × 3; therefore, each greyscale image is repeated three times (for three input channels).

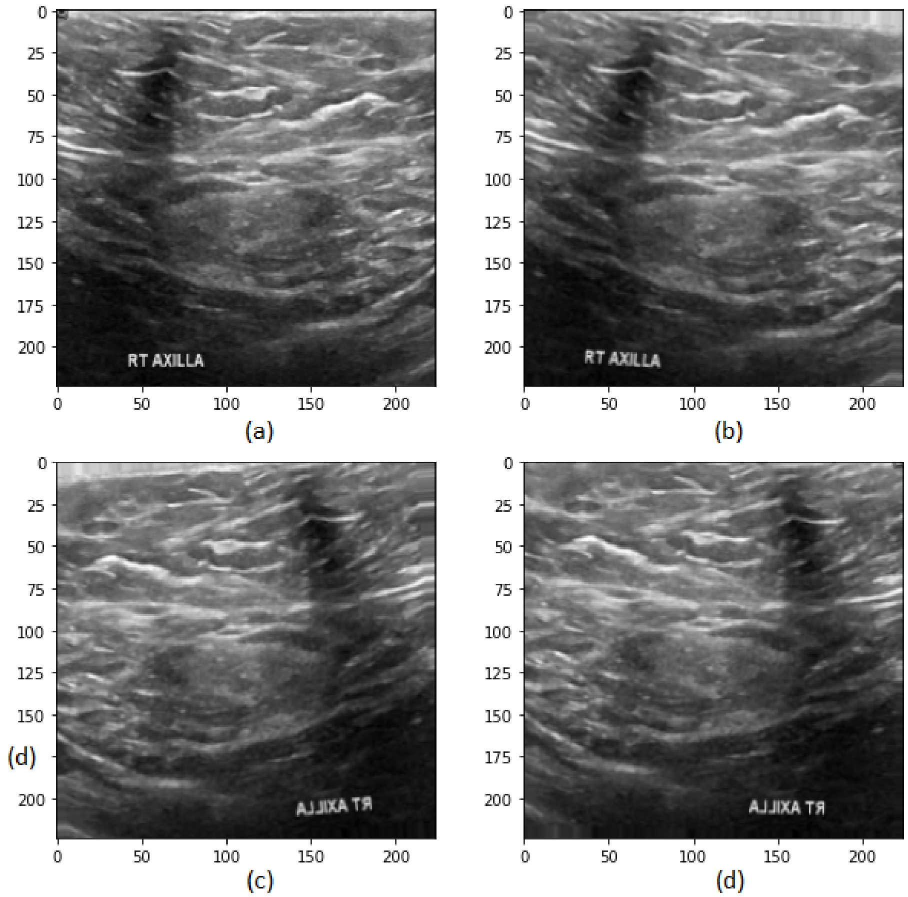


Figure 2.5: (a) Original image; (b)-(d) three generated images by data augmentation generator

In the second experiment, our objective is to use the mask as one of the input images, so we inject two input images with the related mask in the third channel as the input of the model. Therefore we only use the first dataset that provided the masks of the images as well, which is the proper choice for feeding to the model 2.7.

2.2.3 Modifying ResNet-50 For Segmentation

An activation function is a method that is applied to an artificial neural network (ANN) to aid it in learning complicated patterns in data. The function is in charge of determining what should be fired to the next node after the process is finished. An activation function in an ANN does the same thing, and it converts the output signal from the previous cell into a format that may be used by the next cell. The ResNet-50 activation layer matrix weights size is $7 \times 7 \times 2048$ while the output mask size is $224 \times 224 \times 1$. In order to have the same matrix weights size as the image size, we use an upsampler with a rate of 32, which we called direct upsampling. Another way to perform this resizing of weights is to use a decoder after the activation layer. The decoder is consists of six steps as illustrated as boxes in Figure 2.8. Each box in this model includes 2D-Convolutions to train the model while applying upsampling, ReLU activation, 2D-Upsampling, and Batch Normalization. One of the key features of this proposed method is that it continues learning while it is upsampling the masks to the desired size.

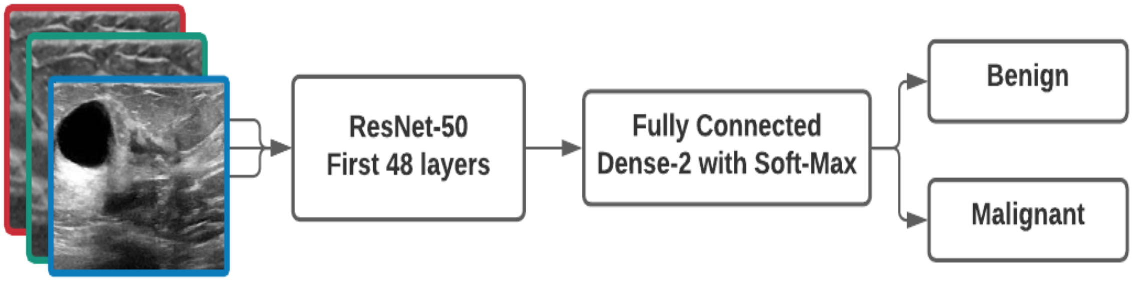


Figure 2.6: A diagram of the proposed single task network to classify benign and malignant; The input images include three greyscale image

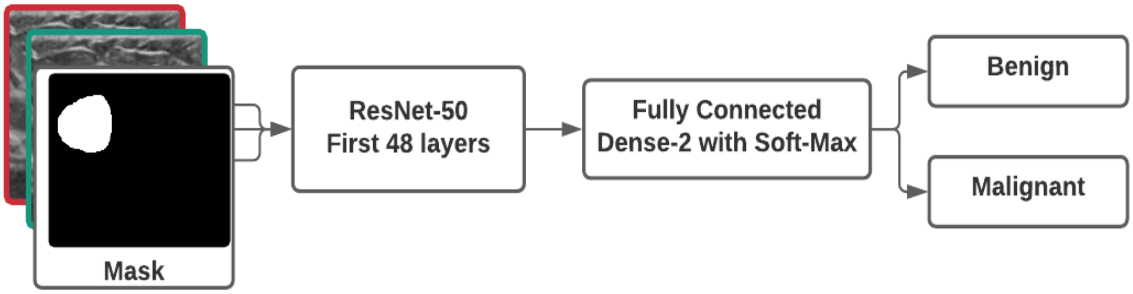


Figure 2.7: STL network to classify benign and malignant with input mask; the input images included two greyscale images and one mask

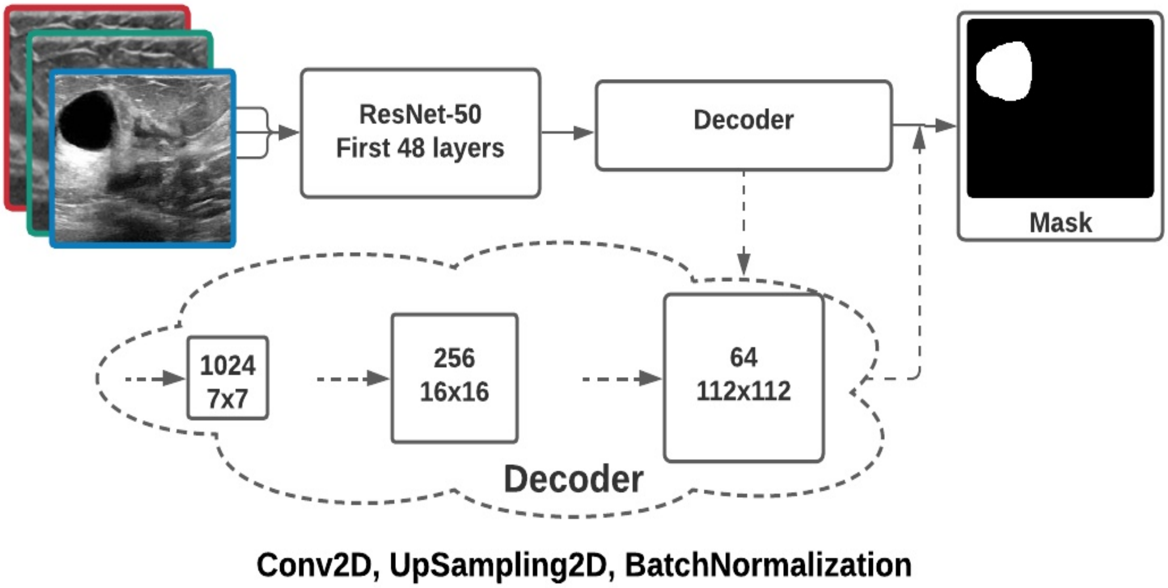


Figure 2.8: The segmentation network based on the ResNet-50

2.2.4 Hyperparameters Tuning

Hyperparameters are the parameters that determine the model architecture, and hyperparameter tuning is the process of finding the perfect model architecture. When building a machine learning model, we will be given design options for defining the model hyperparameters. We often do not know what the best hyperparameters set is for a specific model right away; thus, we would like to be able to experiment with a variety of options.

Optimizer:

Optimizers are techniques or strategies for minimizing the error function (loss function) or increasing production efficiency. Optimizers are mathematical functions that are based on the learnable parameters of the model, such as weights and biases (the bias value allows the activation function to be shifted to the left or right to better fit the data). Optimizers aid in identifying how to change a neural network's weights and learning rate to minimize losses.

For all of our networks, we use Adaptive Moment Estimation (Adam) Kingma and Ba (2014) as the optimizer. The Adam optimizer is a well-known and widely used gradient descent optimization technique. It is a technique for determining adaptive learning rates for each parameter. Like Adadelta and RMSprop, Adam first calculates the square of the gradients v_t and then preserves the exponentially decaying average calculation value. It also maintains an exponentially decaying average of previous gradients mt, similar to momentum. While momentum may quickly go up or down a slope, Adam moves more slowly, thus choosing flat minima in the error levelHeusel, Ramsauer, Unterthiner, Nessler, and Hochreiter (2017). The decaying averages m_t of past and past squared gradients v_t are calculated as follows: The initial moment of the gradients is estimated by mt (the first one), whereas the uncentered variance is estimated by vt (the second one). Adam discovered that m_t and v_t are desired around zero since they are initialized as vectors of 0's, particularly during the early time steps and when the decay rates are low (i.e., $beta_1$ and $beta_2$ are close to 1). It can be compensated for these biases by computing bias-corrected first and second moment estimations:

$$\begin{aligned}\hat{m}_t &= \frac{m_t}{1-\beta_1^t} \\ \hat{v}_t &= \frac{v_t}{1-\beta_2^t}\end{aligned} \tag{1}$$

These values can be utilized to update the parameters, similar to Adadelta and RMSprop, resulting in the Adam update rule:

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{\hat{v}_t} + \epsilon}\hat{m}_t \tag{2}$$

The authors of Heusel et al. (2017) suggest default values of $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^-8$. They show that Adam practically outperforms the other adaptive learning-method algorithms in performance.

Learning Rate:

The stochastic gradient descent optimization approach is used to train deep learning neural networks. The learning rate is a hyperparameter that governs how much the model changes each time the model weights are modified in response to the predicted error. The learning rate controls how quickly the model is adapted to the problem. A too little number may result in a protracted training process that becomes stuck, whereas a too big value may result in learning a sub-optimal set of weights too quickly or an unstable training process. When constructing a neural network, the learning rate may be the most crucial hyperparameter. As a result, it is critical to understand how to examine the impacts of the learning rate on model performance and to develop an understanding of the learning rate's dynamics on model behavior.

Metric Function:

Every machine learning pipeline has performance measurements. They make progress and put a number on it. All machine learning models need a metric to monitor and measure the performance of a model during training and testing. For classification and segmentation, we employed respectively Area Under Curve (AUC) Bradley (1997) and Dice score Sudre, Li, Vercauteren, Ourselin, and Cardoso (2017) as the performance evaluation metric. One of the most extensively used measures for evaluating binary classification algorithms is AUC. A classifier's AUC is the likelihood that a randomly given positive example will be ranked higher than a randomly picked negative example. Let us first define two concepts before moving on to AUC:

- True Positive Rate (TPR) or Sensitivity :

  $\frac{TP}{FN+TP}$ With regard to all positive data points, sensitivity is the fraction of positive data points that are accurately counted as positive (TP).

- True Negative Rate (TNR) or Specificity:

  $\frac{TP}{FP+TN}$ The fraction of negative data points that are accurately classified as negative (FP) out of all negative data points is known as specificity.

- False Positive Rate (FPR):

  $\frac{FP}{FP+TN}$ In comparison to all negative data points, FPR is the fraction of negative data points that are wrongly considered positive.

The values of FPR and TPR are in the range of $[0, 1]$. AUC is the area under the curve of plot FPR vs. TPR at different points in $[0, 1]$.

Using the area under a receiver operating characteristic (ROC) curve, we can plot AUC, a single scalar metric that assesses a binary classifier's overall performance. The AUC value falls between $[0.5, 1]$, with the lowest value representing the performance of a random classifier and the highest value representing the performance of a perfect classifier.

Because it is calculated using the whole ROC curve and includes all possible classification levels, the AUC provides a reliable overall metric for evaluating the effectiveness of score classifiers. The AUC is commonly derived by multiplying the ROC curve by the number of trapezoid regions below it.

The Dice score is frequently used to assess the performance of image segmentation algorithms. (Eq. 3).

$$DiceScore = \frac{2.|A \cap B|}{|A| + |B|} \tag{3}$$

Where A and B are the binary vectors with 1 for elements inside a group and 0 for otherwise, one signifies the ground truth, and the other signifies the classification result.

Loss Function:

Loss functions are used to determine how close an estimated value is to the genuine value. It is a way of determining how effectively a particular algorithm mimics the data. The loss function will return a considerable value if the forecasts are too far from the actual findings. The loss function learns to lower prediction error over time with the aid of some optimization function. For classification and segmentation, we use binary cross-entropy and Dice loss function, respectively.

The binary cross-entropy loss function calculates the loss of an example by computing the following sum (4):

$$binary\ cross\ entropy\ Loss = -\frac{1}{N}\Sigma_{i=1}^{N} y_i . \log(p(y_i)) + (1 - y_i) . \log(1 - p(y_i)) \tag{4}$$

This loss is an excellent measure of how distinguishable two discrete probability distributions are from each other. In this context, y_i is the probability that event i occurs (1 for benign and 0 for malignant), and the sum of all y_i is 1, meaning that precisely one event may occur. Where $p(y)$ is the probability of the point predicted benign for all N points. It adds $log(p(y))$ to the loss for each benign $(y = 1)$, which is the log probability of being benign. Conversely, for each $(y = 0)$, it adds $log(1 - p(y))$, which is the log probability of being malignant, . Finally, by putting a minus sign, we make sure that the loss decreases when the distributions get closer to each other.

To calculate the loss function for the segmentation, we modified the Dice metric, which was explained in the previous section. The Dice loss function is calculated as 1 minus the loss value as in Equ. 5.

$$Dice\ Loss = 1 - \frac{2.|A \cap B|}{|A| + |B|} \tag{5}$$

2.3 Results

Note: for implementing all networks in this thesis, we use Python and "Keras", which is an open-source neural network library Gulli and Pal (2017).

2.3.1 Classification

In order to train and test our classifier model explained in Section 2.2.2, we use both datasets as Al-Dhabyani et al. (2020) and Rodrigues (2017). For this model, the metric is set to AUC, and binary cross-entropy is applied as the loss function. Finally, we adjust the learning rate to 0.0001 for the Adam optimizer.

Figure 2.9 shows a good fit learning curves that train loss and validation loss to decrease to the point of stability with a minimal gap between the two final loss values. This ensures that the overfitting problem was well controlled, and the final model performs on the validation set as well it performs on the train set. To calculate AUC and plot the ROC curve, we use Scikit-Learn Pedregosa et al. (2011) package. As Fig. 2.10 shows, the AUC score is 96.71%.

The objective of the second practice is to apply the mask as one of the input channels. In order to feed the images to the three input channels, we duplicated the original image for the first two channels and then used the mask for the last channel. The input of the networks in this practice needs a mask. Therefore the first dataset is used, which includes the related masks for benign and malignant. For tuning the hyperparameters, we use the AUC metric, binary cross-entropy as the loss function, and Adam optimizer the same as the first practice. For the optimizer, we set the learning rate on 0.0001. Fig. 2.11 shows the learning curves of loss of train and validation. Both curves drop continuously to a point of stability. Fig. 2.12 represents 99.69% for AUC score.

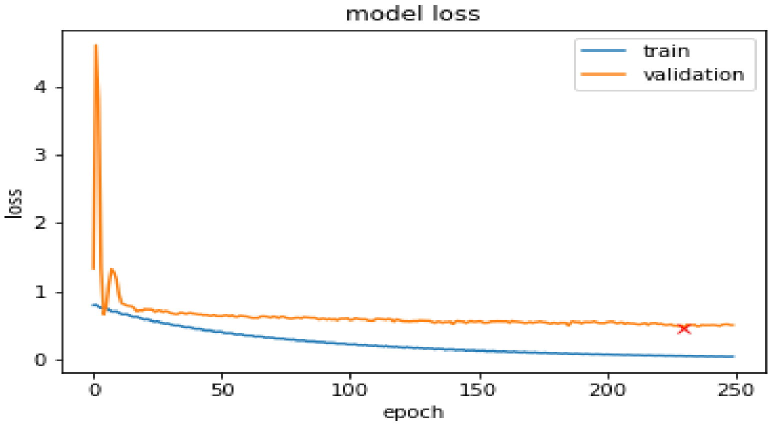


Figure 2.9: First practice: loss of train and validation, using both datasets with modified ResNet-50 as the STL network

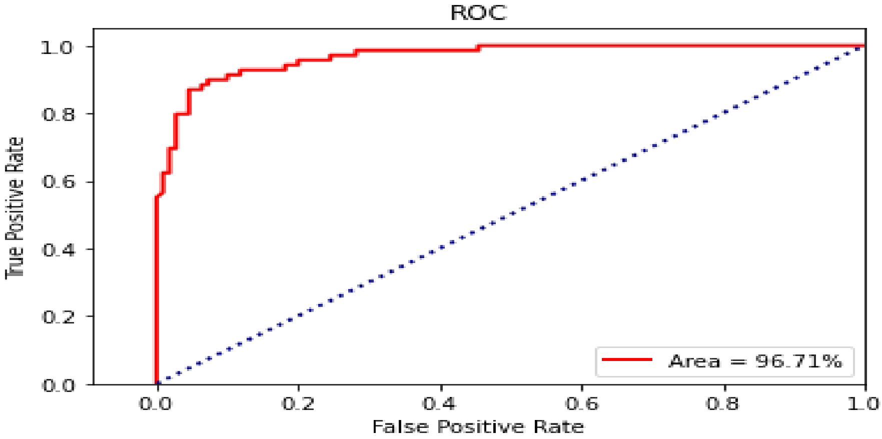


Figure 2.10: First practice: ROC of classification network with original image for the three input channels

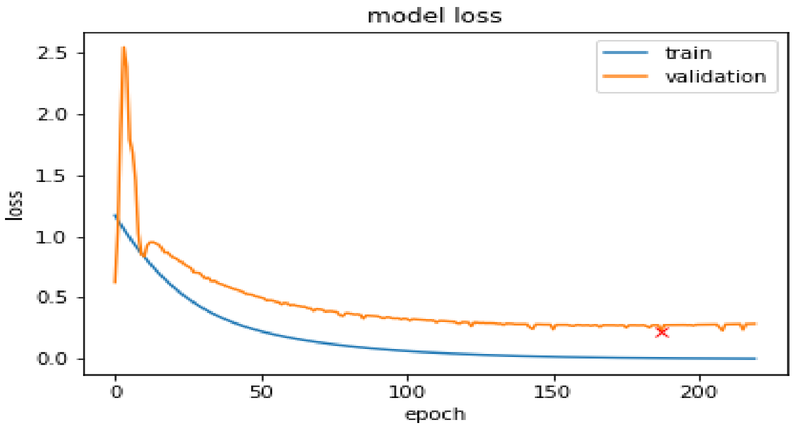


Figure 2.11: Second practice: loss of train and validation, using first dataset (includes mask) with modified ResNet-50 as the STL network

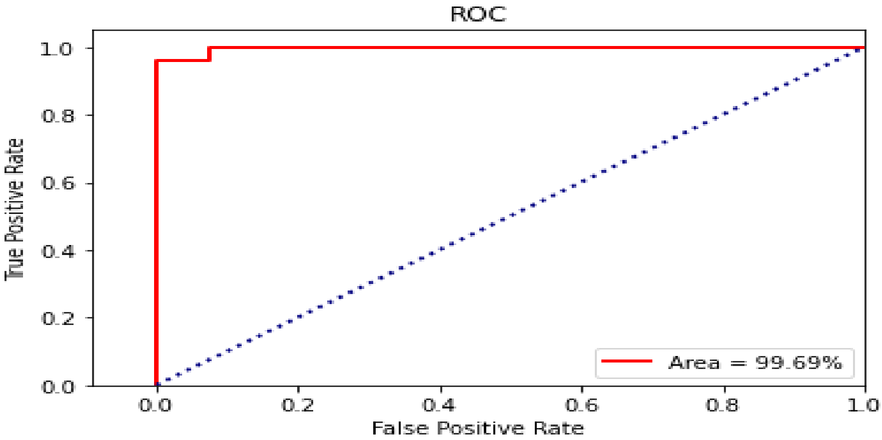


Figure 2.12: Second practice: ROC of classification network with mask as one channel and original image for two channels

2.3.2 Segmentation

In this section, we are going to cover the result of our designed segmentation model based on ResNet-50 and U-Net. For this task of segmentation, we only used the first dataset, which includes masks. For tuning the hyperparameters, Dice is applied for both loss function and metric as well. Then, Adam, with a learning rate of 0.0001, optimizes the weights. By comparing train loss and validation loss in Fig. 2.13, we ensure that overfitting and underfitting do not happen.

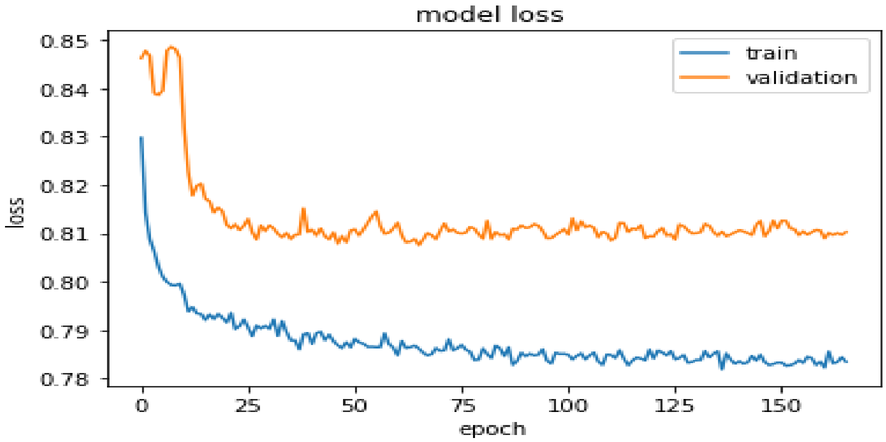
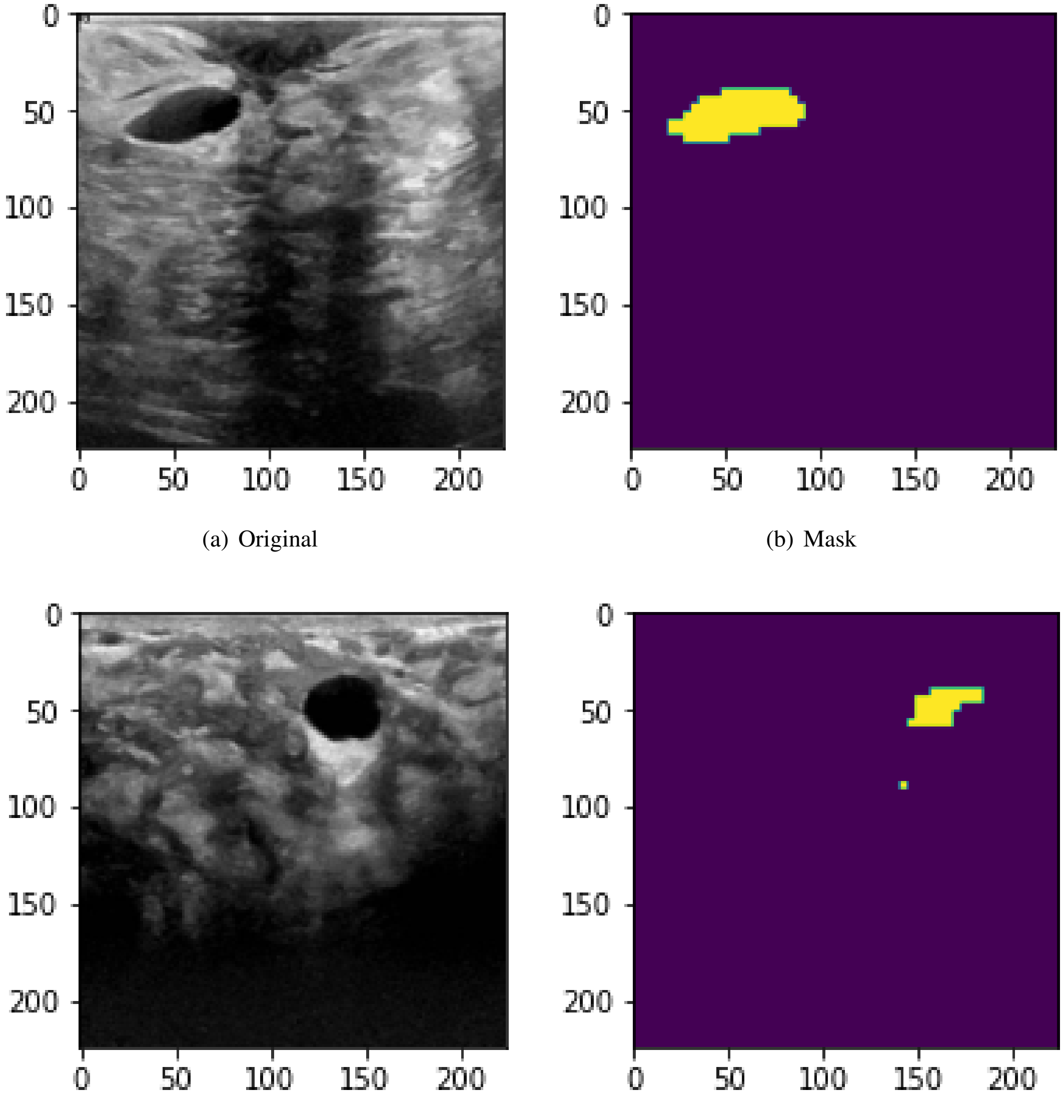


Figure 2.13: Train and validation loss for the segmentation network

To measure how well the model generates the masks, we use the Dice score as the metric, then compare the ground truth mask and the produced mask. As a result, the Dice score is 67.93%. In comparing the ground truth mask with the predicted mask, we mainly find different results. Here in Fig. 2.14(b) illustrates a relatively good segmented mask while Fig. 2.14(d) shows a poor prediction.



(a) Original

(b) Mask

(c) Original

(d) Mask

Figure 2.14: (a) and (c) Original images, (b) well prediction, (d) poor prediction

2.4 Conclusion

In this chapter, we briefly reviewed deep learning classification and segmentation models. Then hyperparameter turning is fully covered, which controls the learning process. Furthermore, we also proposed two networks for classification and segmentation tasks in three phases.

We build a network that predicts benign and malignant images using the ResNet-50 model in the first phase. This network can classify many images by having an AUC score of 96.71%.

In the second phase, we expected a high AUC score with the identical network and used the mask as one of the input channels. So, the result confirms our expectation due to reaching AUC to 99.69%.One disadvantage of the proposed strategy is that, even if our classification model's AUC is relatively high in this phase, we must supply mask as one of the inputs for each prediction. It could become an issue because masks are not available in many datasets. Therefore as future work, we suggest utilizing the MTL model in Chapter 3 to tackle this problem.

Finally, we proposed a learning framework for image segmentation based on the ResNet-50 and U-Net models in the third phase. The task used different steps of upsampling to predict the mask. On the one side, we achieved a Dice score of 67.93% in our segmentation implementation. On the other side, we investigated the predicted masks visually. As a result, our observation showed different achievements, for some of the masks generated well while others looked inappropriate.

Chapter 3

Multi-Task Learning

3.1 Introduction

Multi-task learning (MTL), as one of the emerging techniques in a variety of machine learning applications such as natural language processing Collobert and Weston (2008), computer vision K. He, Gkioxari, Dollár, Girshick, and R-CNN (2017), speech recognition L. Deng, Hinton, and Kingsbury (2013) have been inspired by human learning. More specifically, humans often learn new tasks by leveraging learning the related tasks. Similarly, in the MTL technique, in contrast to STL networks, the network is trained while optimizing more than one loss function (i.e., task). In other words, the network is trained to solve multiple tasks simultaneously. This will consequently expedite the computations during inference time, improve the predictions, and reduce the training time Standley et al. (2020). It further helps to the generalization of the network by exchanging the representations between similar tasks and leads to less risk of overfitting Ruder (2017). However, as explained by Standley et al. (2020), it is crucial to integrate tasks that are compatible with each other.

In ultrasound (US) studies, MTL has recently increased researchers' attention. For example, Behboodi, Rasaee, Tehrani, and Rivaz (2021) used a multi task classification of breast ultrasound image to classify invasive ductal carcinomas (IDC), cysts (CYST), and fibroadenomas (FA). They demonstrated that the value of multi-task learning is improving IDC identification in breast US images. They also observed that increasing the number of classes in deep learning networks boosted their performance. Finally, they proposed a unique technique for adding a backdrop class to ultrasound images. Another similar study Lin et al. (2019), proposed the usage of MTL in the detection of the fetal head standard plane by adopting a classification module alongside a frame detection module based on Fast-RCNN network Girshick (2015). They classified the ultrasonic plane of the fetal head to standard and non-standard images based on the key anatomical structure that appeared in the image. In Ke et al. (2021), they investigated the challenge of accurately localizing object contours from coarse labels in a data-driven context, particularly for weakly contrasted images or objects with complicated borders. They developed a segmentation task alongside a recursive approximation task for partial object region learning. They achieved segmentation masks enhancement in fetal head ultrasound images.

The studies mentioned above were utilized for either segmentation tasks or classification tasks. The idea of MTL for jointly training segmentation and classification in ultrasound images has already been introduced. For example, P. Wang, Patel, and Hacihaliloglu (2018) explored the simultaneous training of segmentation and classification branches of a network for bone surfaces of ultrasound images. They have shown improvements in segmentation masks after adding the classification branch. Similarly, Xie, Shi, Niu, and Tang (2018) proposed a two-stage multi-task network by adopting ResNet and Mask R-CNN networks. However, they evaluated their suggested technique using a private dataset, making it challenging to utilize as a standard network for new datasets. Singh et al. (2019) also proposed a generative adversarial network (GAN) for segmentation and classification of breast ultrasound images. They used segmentation and classification branches as their generator and discriminator networks, respectively. However, training GAN-based networks have always been difficult for new datasets. Therefore, this chapter proposes a novel MTL-based technique for simultaneous training of segmentation and classification tasks for a publicly available breast dataset. Furthermore, in the proposed technique, we take advantage of the existing masks in order to further enhance the predictions. The contributions of our proposed technique can be summarized as:

- MTL-based network for simultaneous segmentation and classification training

- Enhancing performance in both segmentation and classification tasks compared to STL

3.2 Method

In Chapter 2 we came up with two novel algorithms that utilized ResNet-50 K. He et al. (2016) model to classify benign and malignant US images at the first experiment and then generated related segmentation masks by adding a U-Net-based Ronneberger et al. (2015) decoder model to ResNet-50 encoder. Here in this Chapter 3, we propose an MTL network inspired by our findings from Chapter 2 where we integrate classification and segmentation networks to perform both tasks simultaneously. To this end, we employ a U-Net-based encoder-decoder architecture and modify its encoder and decoder in such a way to be able to perform MTL. We further improve our proposed MTL network by taking advantage of the existing segmentation masks. More detailed information is provided in the following sections.

3.2.1 MTL Network Design

As we mentioned earlier, we utilize a U-Net-based network for our proposed MTL network, consisting of three main branches: an encoder, a decoder, and a classifier. The diagram of our proposed MTL network is shown in Fig. 3.1.

Encoder Design Based on the high performance of ResNet-50 in Chapter 2, we set the encoder branch ResNet-50 where it takes inputs with the size of $224 \times 224 \times 3$. We take the output of two different layers to be fed to decoder and classifier branches. The output of the 48th layer with the size of $7 \times 7 \times 2048$ is fed to the decoder branch, while the output of the average-pooling layer of ResNet-50 is fed to the classifier. Similar to what we explained in Section 2.2.2 each greyscale image is repeated three times to meet the input size.

Decoder Design The decoder consists of six blocks, each made of a convolution layer with ReLU activation, followed by an upsampling layer with a kernel size of 2×2, and a batch normalization layer. The output of the decoder branch has a size of $224 \times 224 \times 1$ in order to have the same size as the input image size, and during training, the step is optimized with ground-truth masks. We refer to the encoder and decoder branches as our segmentation branches for simplicity.

Classifier Design The output of the average-pooling layer of ResNet-50 with the size of 1 × 1 × 2048 is fed a fully-connected layer with only two nodes as we have the binary classification problem (i.e., benign versus malignant). Finally, a Softmax activation layer provides the probabilities for each class.

3.2.2 Dataset

The dataset that is used is the first dataset that we explained earlier in Section 2.2.1. It includes 780 US images with their corresponding segmentation masks. However, in the current chapter, we only used benign and malignant images leading to a total of 647 US images.

3.2.3 Experiments

For training segmentation and classification branches of our MTL network, we used similar hyperparameters as explained in Section 2.2.4. Furthermore, as we have two branches, the loss function integrates the segmentation and classification losses defined as Dice similarity loss and

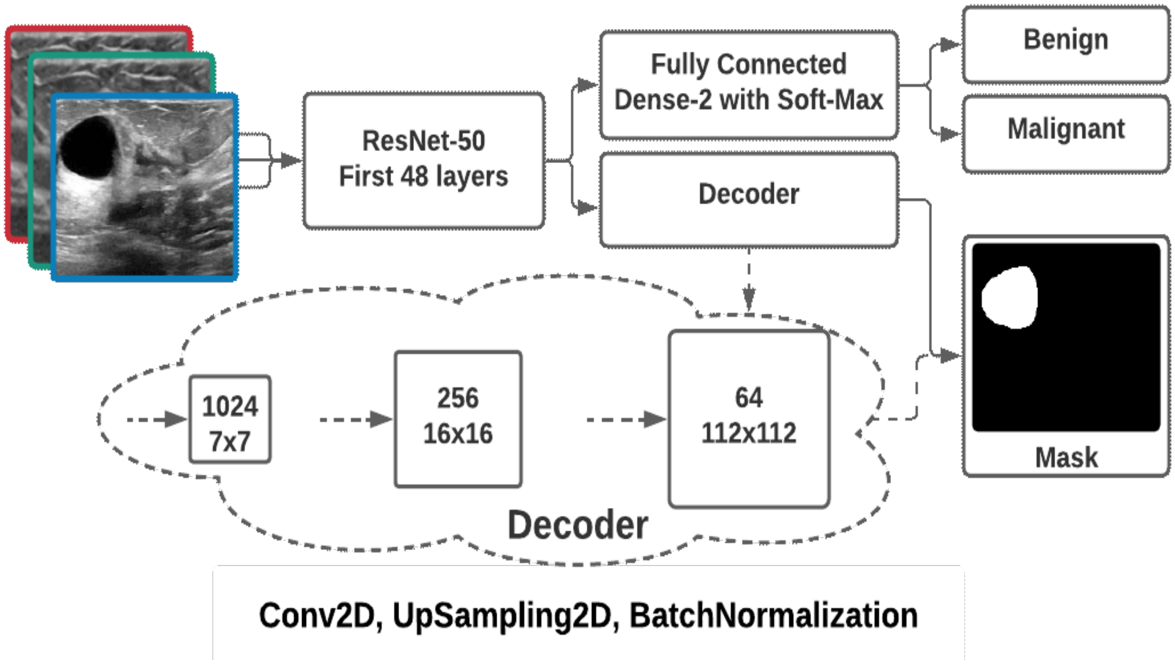


Figure 3.1: A diagram of the proposed MTL network, showing the top branch for classification and the bottom branch (decoder) for segmentation.

Table 3.1: Hyperparameters tuning

Branch	*Classification*	*Segmentation*
Loss function	*BCE loss*	*Dice loss*
Evaluation Metric	*AUC*	*Dice score*
Learning rate	*0.0001*	*0.0001*
Optimizer	*Adam*	*Adam*

cross-entropy loss, respectively (see Eq. 5 and 4 of Chapter 2). Consequently, the final loss is defined as:

$$loss_{MTL} = \alpha \, . \, loss_{\,binary\, cross\, entropy} + \beta \, . \, loss_{Dice} \tag{6}$$

where α and β are the coefficients of binary cross entropy loss and Dice loss, respectively. By tuning these coefficients we can regulate the focus of the network such that it can give more impact to either segmentation or classification tasks. More details will be provided in Section 3.3. For training, we make sure to save the best model based on validation loss.

3.2.4 Evaluation Metrics

Similarly to Chapter 2, we use Dice similarity score (see Eq. 3) to evaluate predicted segmentation masks and AUC scores to evaluate classification performance. For more details on Dice similarity and AUC scores, please refer to Section 2.2.4.

3.3 Results

As we explained earlier, our experiments are carried out on the first dataset that incorporates segmentation masks and classification labels. By tweaking the α and β of MTL loss function in Equ. 6, we observe the better performance of our MTL network for both segmentation and classification branches compared to what we observed in experimenting segmentation network and classification network separately in Chapter 2 (i.e., in STL). To be more clear, by setting α and β to 0.1 and 1, respectively, the segmentation performance of our MTL network has improved to

Table 3.2: Multi Task Learning Results

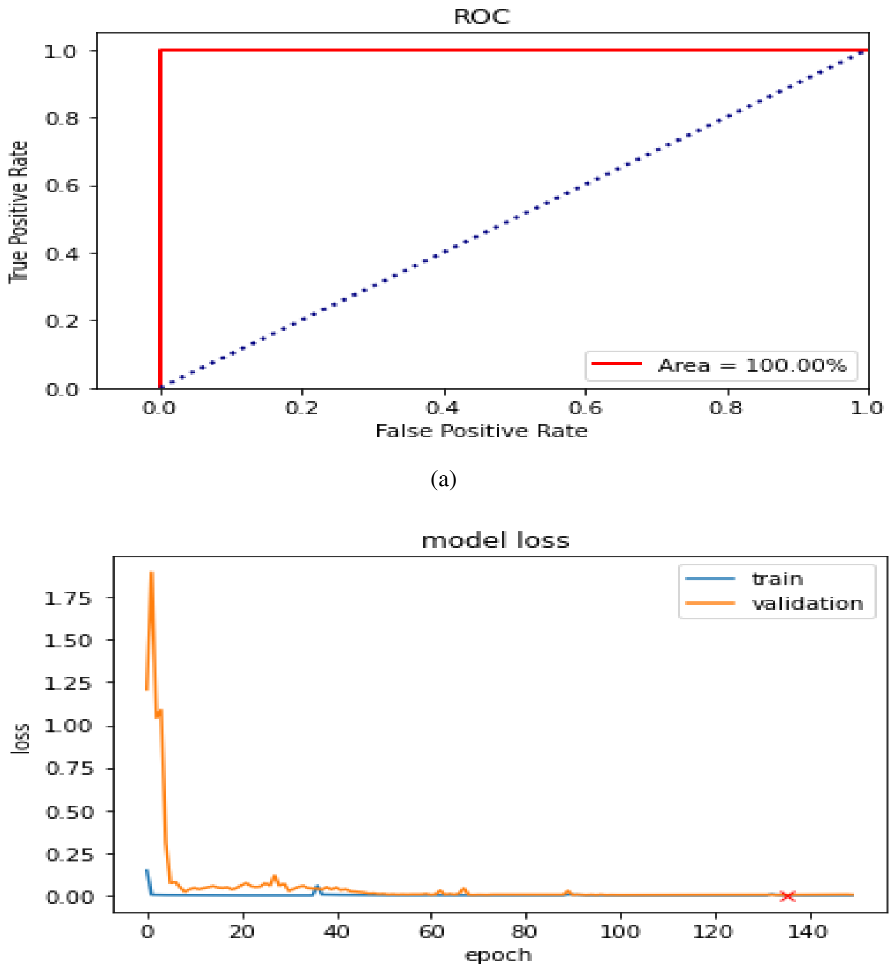
Loss Function Weight		***Result***		
α	β	**Dice**	**AUC**	**Useful for**
0.01	*10*	*41.67%*	*100%*	*Classification*
0.1	*1*	*88.19%*	*90.48%*	*Segmentation*

Table 3.3: Comparison of our STL and MTL networks.

Dice Score	
STL segmentation network	***MTL segmentation branch***
67.93%	*88.19%*
AUC Score	
STL classification network	***MTL classification branch***
96.71%	*100%*

the Dice score of 88.19%, however; based on our results obtained in Chapter 2, the Dice score for the segmentation network in STL was 67.93%. Similar behavior has been achieved for the classification network. By setting α and β to 0.01 and 10, respectively, the classification AUC score has been improved from 96.71% in STL to 100% in MTL. Table 3.3 presenting the comparison of our MTL and STL networks. As a result, combining the classification and segmentation branches yields superior results compared to STL. Table 3.2 summarizes our MTL results. This table shows that the segmentation quality is measured by two metrics, Dice similarity, and AUC scores. The Dice metric represents how closely the predicted area of each particular lesion instance matches the one in the ground truth image. The AUC metric measures the ability of the classifier branch to distinguish between benign and malignant. The AUC score for the classifier branch surprisingly hits 100%, which proves that the network can classify all benign and malignant images perfectly.

Figure 3.2(a) illustrates the ROC curve for predicted class label probabilities. Moreover, Fig. 3.2(b) represents train and validation loss during the training step when setting the focus of the MTL network to focus more on the classification branch rather than the segmentation branch; consequently, it can be confirmed that no overfitting and underfitting has happened. Even though the

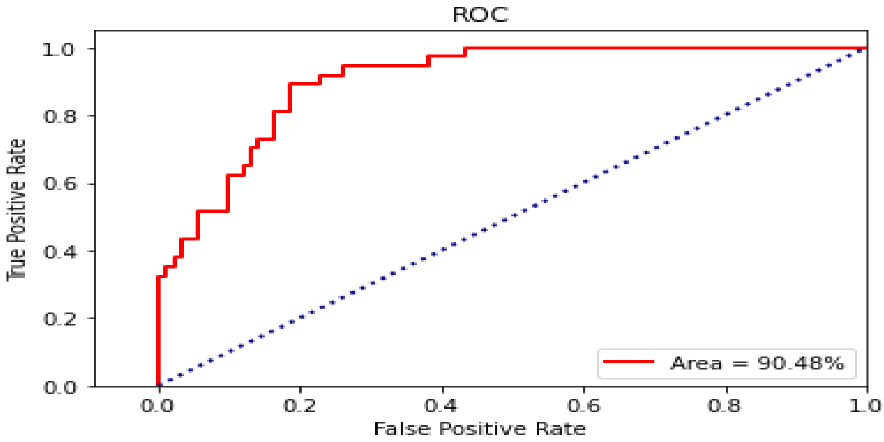


(a)

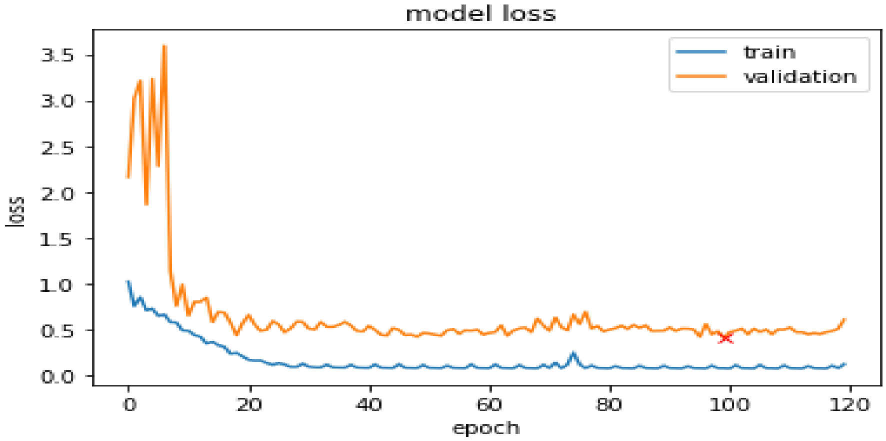
(b)

Figure 3.2: MTL classification branch: (a) ROC curve, (b) Train and validation loss

Dice score for the segmentation branch (which is not our objective) is 41.67%, which is quite low but it helped the classification branch to predict class labels more accurately. Similarly, if we set the focus of our MTL network on the segmentation branch, the training and validation loss shows no overfitting and underfitting during the training step as shown in Fig. 3.3(b). The ROC curve is also shown in Fig. 3.3(a). Comparing the results of the MTL segmentation branch with the STL segmentation network, we confirm that having the classification branch can boost the performance of the network towards wisely segmentation and reaching the Dice score of 88.19%.



(a)



(b)

Figure 3.3: MTL segmentation branch: (a) ROC curve, (b) Train and validation loss

3.4 Conclusions

We have demonstrated that a combination of classification and segmentation in the network with proper hyperparameters does significant effects on the predictions of either classification or segmentation. On the on hand, we primarily aimed to use MTL to create a network to classify benign and malignant images based on the ResNet-50 model. AUC for STL applied on both datasets is 96.71% and for the same model with the mask as one of the input channels is 99.69% then finally for MTL is 100%. As indicated in Table 3.3 it could enhance accuracy by 3.29% by utilizing MTL.

On the other hand, our second goal, which entirely is segmentation improvement, the MTL helped the network to dramatically increase the Dice score 20.26% (from 67.93% to 88.19%) as shown in Table 3.3.

Chapter 4

Explainable AI

4.1 Introduction

In previous chapters, we discussed the applications of convolutional neural networks (CNN) in ultrasound imaging, notably in classification and segmentation tasks. Despite the superior performance of these techniques in numerous applications, they are imperfect solutions to real-life medical researches. These techniques are hard to interpret. Thus when a network either succeeds or fails, the user wonders about intuitive and understandable reasons for both successes and failures Lipton (2018). Interpreting a network's successes and failures is one of the top and yet unsolved challenges in CNN-based techniques as it helps in conveying helpful information about the network. The capacity to explain and interpret a network helps recognize network failure modes as well as create adequate network reliability Selvaraju et al. (2017). Several methods for interpretability of models have been proposed Dosovitskiy and Brox (2016); Koh and Liang (2017); Lundberg and Lee (2017); Selvaraju et al. (2017); Simonyan, Vedaldi, and Zisserman (2014); Zeiler and Fergus (2014); Q. Zhang, Wu, and Zhu (2018).

In the field of image recognition and classification, gradient-weighted class activation maps (GRAD-CAM) proposed by Selvaraju et al. (2017) have been widely used for interpreting classification networks. Grad-CAM is a type of post-hoc attention for creating heatmaps to highlight class-specific regions of images from an already-trained network. It offers informative visualization maps for increasing the transparency of the network. This method employs the gradients of any target class to produce a coarse localization map that emphasizes the critical area in the image. To be more specific, the heat maps produced by GRAD-CAM illustrate where the network is looking at for the given image. GRAD-CAM heat maps have also been applied as a standard visualization technique for clarifying disease detection in medical images. For example, Jin et al. (2020); Oh, Park, and Ye (2020); Panwar et al. (2020); Rajpal, Lakhyani, Singh, Kohli, and Kumar (2021); Umair et al. (2021) utilised GRAD-CAM based color visualization approach in either X-ray or computed tomography (CT) images providing better understanding for detection of COVID-19 cases. Many researchers such as C.-T. Cheng et al. (2019); Nguyen et al. (2019); Rouhafzay et al. (2020); Sánchez Fernández et al. (2020) used GARD-CAM visualization in magnetic resonance images (MRI) for detection of breast lesions, tubers in tuberous sclerosis complex, hip fractures, and uveal melanoma, respectively.

In ultrasound (US) imaging, van Sloun and Demi (2019) developed a weakly supervised CNN-based algorithm in removing B-line artifacts from lung US images. They leveraged GRAD-CAM in performing B-line localization directly from activation maps prior to the denoising step. In another study, T. He et al. (2019) employed GRAD-CAM alongside to multi-layer perceptron (MLP) for crucial variable extraction in patients with lung cancer. Similarly, Dastider, Sadik, and Fattah (2021) deployed GRAD-CAM as a visualization technique to show the attention maps of their proposed classification method. For skin US image classification, Czajkowska, Badura, Korzekwa, Płatkowska-Szczerek, and Słowińska (2021) integrated GRAD-CAM heat maps to their proposed classification network for further measuring the reliability of their method.

As discussed in previous chapters, the main target of the current thesis work is exploring breast US images in both classification and segmentation tasks. Thus, we further investigate the interpretation of our classification network proposed in Chapter 3 similarly to Eskandari, Du, and AlZoubi (2021); Habib et al. (2020); Misra et al. (2021); Rodríguez-Salas, Seuret, Vesal, and Maier (2021). As mentioned above, GRAD-CAM is a measure to interpret better and explain the network behavior. To this end, in this chapter, we illustrate the GRAD-CAM heat maps to comprehend the performance of our proposed classification model and provide human-understandable justifications to the decisions.

4.2 Method

In this section, first, we provide more details about Class Activation Mapping (CAM) introduced by B. Zhou, Khosla, Lapedriza, Oliva, and Torralba (2016) for the first time. Then we explain Gradient-weighted Class Activation Mapping (GRAD-CAM), which is the current chapter's primary objective.

4.2.1 Class Activation Mapping

CAM provides visual explanations for a specific class for a CNN-based classification network. It builds a localization map for the networks that have a global-average-pooling (GAP) convolutional layers before the final Softmax/Sigmoid layer in their design. Given k features maps $A^k \in \mathbb{R}^{u \times v}$ of width u and height v as the input to the GAP layer, then these feature maps are spatially pooled, leading to identical weights w_k. Next, a y^c score for each class c is generated by applying a linear transform on A^k feature maps with their corresponding w_k weights as shown in Eq. 7.

$$y_c \; = \; \Sigma_k \, \omega_k^c \, \frac{1}{Z} \, \Sigma_i \, \Sigma_j \; A_{ij}^k \tag{7}$$

where $Z = u \times v$ is the total number of elements. For visualization reasons, the CAM heat maps are normalized between 0 and 1. Heat maps based on CAM technique are restricted to networks which do not contain any fully-connected layers in their design. Therefore, for networks with fully-connected layers, in order to obtain CAM heat maps, they need to be re-trained by replacing the fully-connected layers with GAP layers.

4.2.2 Gradient-Weighted Class Activation Mapping

Unlike CAM, GRAD-CAMP heat maps are applicable to a significantly broader range of CNN-based networks. The way the feature maps are weighted to create the final heat maps differs between CAM and Grad-CAM. For obtaining GRAD-CAM heat maps in generic CNN-based networks, first the gradient of y^c, the raw output of the network before final application activation (i.e., Softmax/Sigmoid) function, concerning feature mappings A of a convolutional layer (i.e.$\frac{\partial y^c}{\partial A_{ij}^k} \in \mathbb{R}^{u \times v}$)

is calculated to obtain the class-discriminative localization map $L^c_{Grad-CAM} \in \mathbb{R}^{u \times v}$ defined as:

$$L^c_{Grad-CAM} = ReLU\left(\Sigma_k a^c_k A^k\right) \tag{8}$$

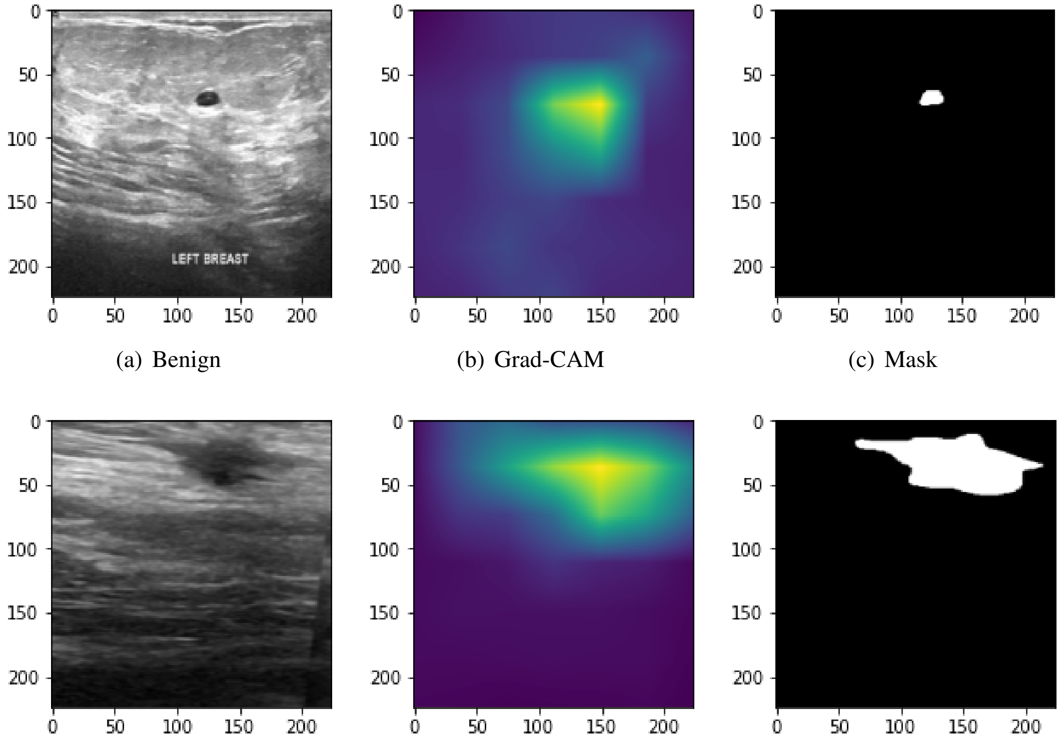
where the weights a^c_k are calculated by

$$a^c_k \ = \ \frac{1}{Z}\,\Sigma_i\,\Sigma_j\,\frac{\partial y^c}{\partial A^k_{ij}}, \tag{9}$$

where $Z = u \times v$ is the total number of elements.

In Eq. 8, a^c_k are the weights that encapsulate the importance of feature map k for a target class c and indicate a partial linearization of the deep network downstream from A. In general, y^c does not have to be a class score; instead, it can be any differentiable activation. The Grad-CAM heatmap is a weighted mixture of feature maps, just like in CAM, but it is followed by a ReLU activation function. As a consequence, a coarse heatmap is created, which is then adjusted for visualization. Aside from the ReLU in Eq. 8, Grad-CAM is a generalization of CAM (ω^c_k are the exact a^c_k where CAM can be applied) to any CNN-based networks design (CNNs with fully-connected layers, ResNets, CNNs stacked with Recurrent Neural Networks (RNNs), and so on).

Here, by aiming above equations, we explain the implementation of Grad-CAM in our STL classification network explained in Chapter 2. There are many layers in deep learning to extract features from an image, and as the model becomes more complex, visual interpretability becomes more important Selvaraju et al. (2017). Only output layer decisions are explained in this thesis. Therefore, in the RestNet-50 model, the activation layer includes the majority of visual information linked to the input image very before the final layers. The activation layer (layer number 48) is a 7×7 matrix with 2048 channels, so we will end up with a weighted matrix that is the same size as the input image by upsampling it 32 times. Instead of upsampling the matrix with a ratio of 32, we can use this approach to include decoder blocks from U-Net to predict the heatmap which is corresponds to the mask area. As a result, we could create our classification model.



(a) Benign

(b) Grad-CAM

(c) Mask

(d) Malignant

(e) Grad-CAM

(f) Mask

Figure 4.1: The original benign image at the left; feature map in the middle; ground truth mask of the original image at right. The original malignant image at the left; feature map in the middle; ground truth mask of the original image at right.

4.3 Results

Figure 4.1 qualitatively illustrates the generated GRAD-CAM heat maps from our STL classification network for the benign and malignant images. The heat maps highlight the predicted weight of the network and give an intuition about which features in the cross-section help judge the class. We can visually verify that visualization heat maps can explain the network by comparing the heat maps and the ground truth segmentation masks. Thresholding these heat maps provides segmentation masks for datasets where we do not initially have the manually prepared ground truth masks. Furthermore, it can be used as an initial point in finding bounding boxes in unsupervised scenarios where the initial points of the bounding boxes are not applicable.

4.4 Conclusions

This chapter presented a gradient-based visualization technique for deep classification convolutional networks. This technique generates an artificial image that represents a class of interest while highlighting the areas of the image that are discriminative concerning the given class. Moreover, we will also use this method in Chapter 5 to visualize the adversarial attacks on the network.

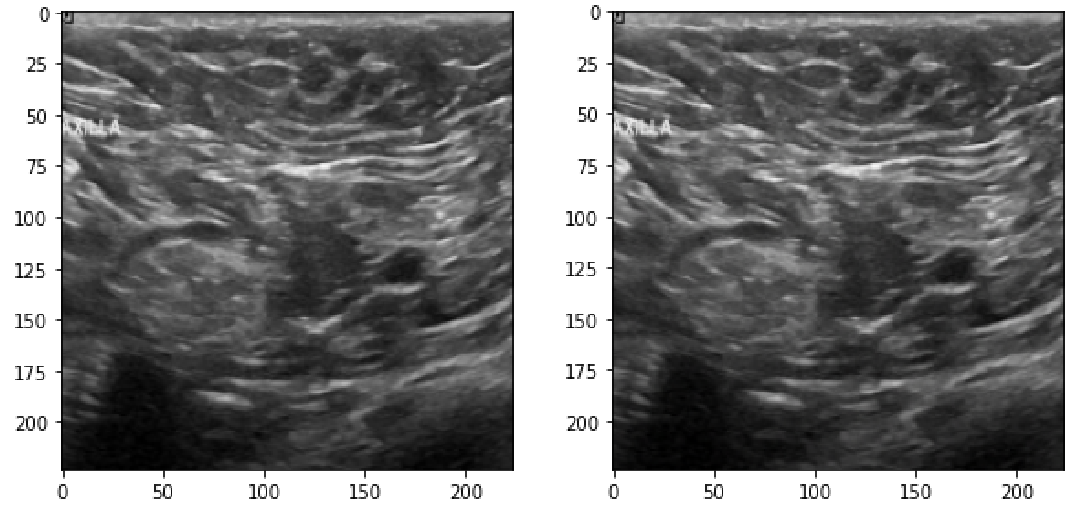
Chapter 5

Susceptibility To Adversarial Attacks

5.1 Introduction

Confidence in a deep learning model is critical, especially in the medical field. Even though explaining a model was made possible by using Grad-CAM as covered in Chapter 4, it still needs to be validated with different data qualities. It is known that ultrasound images can be corrupted by different sources of noise, and their appearance can substantially change by using a different frequency or beamforming approach. These changes can corrupt the classification results or the Grad-CAM. Moreover, the predictions of the deep learning networks are susceptible to adversarial attacks I. J. Goodfellow et al. (2014); Kurakin et al. (2016); Moosavi-Dezfooli et al. (2016). An adversarial attack involves gently altering a source image by adding small artificially crafted perturbations to the image intensities in such a way that the alterations are practically imperceptible to the naked human eye. The modified image is referred to as an adversarial perturbed image that could hinder the accurate decision-making capabilities of a classification network. In other words, the perturbed image is misclassified when presented to the classification network, whereas the original image is correctly classified Das and Rad (2020). Fig. 5.1 shows an example of the original breast US image versus its perturbed version.

Finlayson et al. (2018) examined the feasibility of adversarial attacks even for extremely accurate medical classifiers. They established several experiments to investigate the robustness of CNN-based classifiers to perturbation medical images. They observed that all the models they used



(a) Original

(b) Perturbed

Figure 5.1: The original image on the left; the perturbed image on the right.

were uniquely susceptible to adversarial attacks. Byra et al. (2019) and Byra et al. (2020) have shown that the impact of adversarial attacks during the US image reconstruction steps can easily fool the network and hinder the predictions. More specifically, Byra et al. (2019) applied adversarial attacks on radio-frequency (RF) signals prior to constructing US B-mode images. They showed that even little changes to the breast US image reconstruction algorithm could have a large detrimental influence on classification performance. Similarly, Byra et al. (2020) made minor tweaks to the parameters relating to the reconstruction of liver US images to demonstrate a dramatic drop in the deep CNN-based classification network's classification performance, which resulted in incorrect output. Some researchers apply adversarial attacks on the intensities of the US images Byra et al. (2020, 2019), in contrast, Becker et al. (2019) proposed a methodology based on generative adversarial networks (GAN) to apply perturbation on US images. They tested multiple readers to distinguish perturbed and original US images, and they found that the modified images resulted in a significant decline in readers' performance.

In this chapter, we implement the adversarial attacks on the input images to demonstrate that adversarial examples in breast US images are capable of manipulating CNN-based networks. We, therefore, argue that researchers, especially in clinical US imaging, should be aware of the current vulnerabilities of CNN-based networks and raise research communities' concerns in further investigations of medical learning systems. We employed the "fast gradient sign method" introduced by I. J. Goodfellow et al. (2014) where the noise (not random noise) added to the input image has a direction same as the gradient of the network's cost function concerning the input image.

5.2 Method

In this section, we use the fast gradient sign method (FGSM) to inject noise into the input image, which is a common adversarial perturbation approach and introduced by I. J. Goodfellow et al. (2014) for the first time. As we discussed earlier, the adversarial examples refer to input images (i.e., x) that have weighted noise (i.e., $\epsilon \times \eta$) invisible to the naked human eye. In other words, a network that correctly classifies the input image x is then vulnerable to the input image $x + \epsilon \times \eta$, meaning that it will misclassify it (ϵ controls how much noise to be added). Therefore, in FSGM, the noise, η, is calculated based on the gradient of the loss function with respect to the input image pixels. To be more clear, η is defined as:

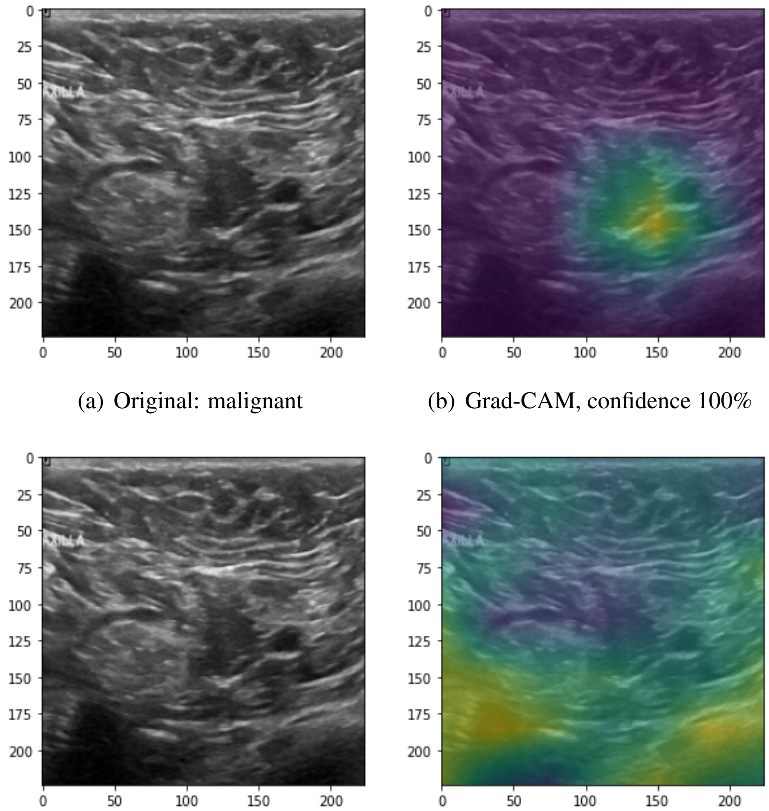
$$\eta = sign(\nabla_x J(\theta, x, y)) \tag{10}$$

where θ is the model's parameter, x is the model's input, y is the predicted label linked to x, J is the trained model's cost function, and $sign(.)$ is the sign function.

It is worth noting that the gradient is just a directional tensor and provides information on which direction to move. To this end, an input image is fed to the trained network through forward-propagation, then the gradients with respect to the input image are calculated through back-propagation. Next, the results are added to the input image. This method employs iterative processes, in which the noise is incrementally applied to the input image. In our experiment, we used $\epsilon = 0.1$ with 25 iterations to ensure that our STL classification network is sufficiently deceived while the noisy input image is visually identical compared to the original image.

5.3 Results

After the disturbance, there are primarily two sorts of impacts on the images. The adversarial perturbation changes the feature map, as seen in Fig. 5.2. As a result, the feature map weights move from the center to the bottom-left of the image. Yet, the predictions on the original and perturbed images are the same with extremely high confidence. The adversarial perturbation on the image deceives the network. Fig. 5.3 indicates that the model categorized the image as malignant with 100% confidence before perturbation, while the model wrongly predicts benign with high confidence after the adversarial perturbation. The difference between the original and perturbed images in Fig. 5.4. The maximum value of the difference is 1, whereas the maximum value of B-mode images is 255. In other words, the changes in the B-mode image caused by the adversarial attack is very small.



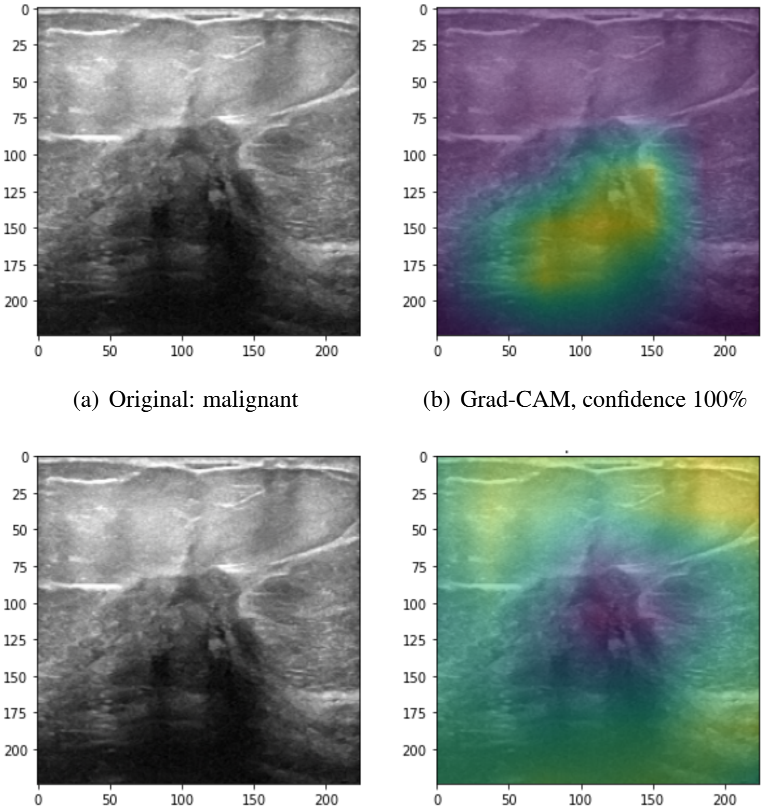
(a) Original: malignant

(b) Grad-CAM, confidence 100%

(c) Perturbed: malignant

(d) Grad-CAM, confidence 99.03%

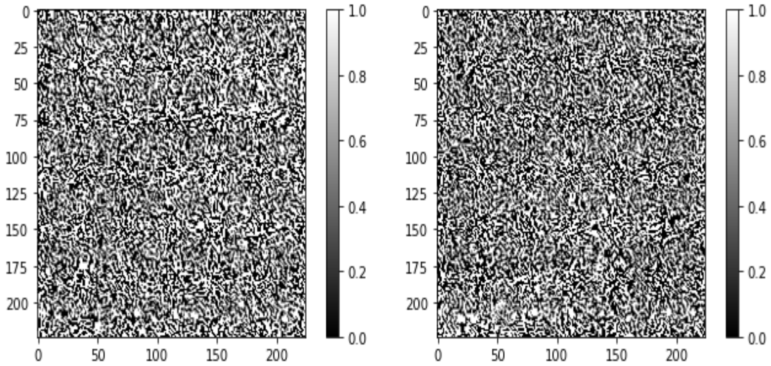
Figure 5.2: The original image on the top left; feature map on the top right; the perturbed image on the bottom left; feature map on the bottom right following adversarial perturbation. The same classification was predicted.



(a) Original: malignant (b) Grad-CAM, confidence 100%

(c) Perturbed: benign (d) Grad-CAM, confidence 96.73%

Figure 5.3: The original image on the top left; feature map on the top right; the perturbed image on the bottom left; feature map on the bottom right following adversarial perturbation. The different classification was predicted.



(a) Different between images Figs. 5.2 (b) Different between images Figs. 5.3

Figure 5.4: The differences between images before and after adversarial attacks of Figs 5.2 and 5.3. Note that the range in these images is between 0 to 1. The range in B-mode intensities is 0 to 255.

5.4 Conclusions

Even though the designed MTL based on the ResNet-50 model can identify the input images as benign or malignant with a high accuracy (100%), the images before and after our modest adversarial perturbations seem virtually the same. While adversarial assaults on CNNs are well-known, our findings demonstrate that interpretations of breast ultrasound images are also subject to similar attacks. As we mentioned, missed diagnosis has been a critical public health concern in clinical diagnosis and treatment, and it may cause disease deterioration and reduce the cure rate. Therefore, we believe that our research is a significant step towards developing reliable deep learning computer-assisted diagnosis systems.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

Chapter 1 discussed the fundamentals of ultrasonic imaging, including a brief discussion of US physics and beamforming in the transducer. In addition to B-mode pictures, which are commonly used to explore the different tissue layers of the human body, US echoes might give a lot more information about the tissue's mechanical and microstructural qualities. Then we had an overview of the deep learning method for classification, segmentation, and multi task learning. Furthermore, we briefly talked about the ways to explain a network and how to affect the network to misclassify by using adversarial attacks.

In Chapter 2, we suggested deep learning techniques, mainly segmentation and classification. Therefore, this chapter started with designing two different convolutional neural networks based on the ResNet-50 network to classify benign and malignant images and then predict a mask for segmentation purposes. After that, we explained related equations with hyperparameters and how they affect network prediction.

In Chapter 3 we elaborated on the Chapter 2 to reach a better solution based on the multi task learning. Again by modifying the ResNet-50 network and the rich experiences from the previous chapter, the new network is designed. As a result, MTL also produced a better result for classification and segmentation.

The goal of Chapter 4 was to find a way to explain the deep neural networks. Therefore we used a class-discriminative localization technique Gradient-weighted Class Activation Mapping (Grad-CAM), to make our model transparent. By applying the method, we explain our network visually.

Adversarial assaults on CNNs are well-known, therefore in the last Chapter 5, we challenged our network by using adversarial attacks. Noise is injected into the input image by applying adversarial attacks. The produced image before and after adversarial perturbations appear to be almost identical. Our experience demonstrates that interpretations of breast ultrasound images are subject to similar attacks.

6.2 Future Work

The work detailed in this thesis can be improved on a number of levels. Here are some research project ideas for the future:

- Because ResNet-50 only accepts a three-channel input, we tripled one channel before applying it to the network. Instead of tripling one input channel, we may use another piece of information for the other channels, such as the histogram equalized of the original picture.

- In Chapter 2 for the segmentation's decoder part, upsamplers are applied to increase the image size. Upsampling causes aliasing; therefore, data loss accrues. In order to raise the resolution of the output mask, we can minimize the upsamples and increase additional 2D-Convolution.

- In Chapter 5, instead of changing ResNet-50 and adding U-Net to the classification model to improve segmentation results, it may be possible to make U-Net the base model and then add the classifier branch to the model.

- In this thesis, we placed adversarial attacks on the input images to misclassify benign and malignant. For future work, we can stress the network by applying the DeepFool Moosavi-Dezfooli et al. (2016) method on the input images. .

References

Al-Dhabyani, W., Gomaa, M., Khaled, H., & Fahmy, A. (2020). Dataset of breast ultrasound images. *Data in brief*, *28*, 104863.

Aldrich, J. E. (2007). Basic physics of ultrasound imaging. *Critical care medicine*, *35*(5), S131–S137.

Antropova, N., Huynh, B. Q., & Giger, M. L. (2017). A deep feature fusion methodology for breast cancer diagnosis demonstrated on three imaging modality datasets. *Medical physics*, *44*(10), 5162-5171.

Becker, A. S., Jendele, L., Skopek, O., Berger, N., Ghafoor, S., Marcon, M., & Konukoglu, E. (2019). Injecting and removing suspicious features in breast imaging with cyclegan: A pilot study of automated adversarial attacks using neural networks on small images. *European journal of radiology*, *120*, 108649.

Becker, A. S., Mueller, M., Stoffel, E., Marcon, M., Ghafoor, S., & Boss, A. (2018). Classification of breast cancer in ultrasound imaging using a generic deep learning analysis software: a pilot study. *The British journal of radiology*, *91*, 20170576.

Behboodi, B., Rasaee, H., Tehrani, A. K., & Rivaz, H. (2021). Deep classification of breast cancer in ultrasound images: more classes, better results with multi-task learning. In *Medical imaging 2021: Ultrasonic imaging and tomography* (Vol. 11602, p. 116020S).

Behboodi, B., & Rivaz, H. (2019). Ultrasound segmentation using u-net: learning from simulated data and testing on real data. In *2019 41st annual international conference of the ieee engineering in medicine and biology society (embc)* (pp. 6628–6631).

Bellman, R. (1966). Dynamic programming. *Science*, *153*(3731), 34–37.

Bradley, A. P. (1997). The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition*, *30*(7), 1145–1159.

Burckhardt, C. B. (1978). Speckle in ultrasound b-mode scans. *IEEE Transactions on Sonics and ultrasonics*, *25*(1), 1–6.

Byra, M., Styczynski, G., Szmigielski, C., Kalinowski, P., Michalowski, L., Paluszkiewicz, R., … Nowicki, A. (2020). Adversarial attacks on deep learning models for fatty liver disease classification by modification of ultrasound image reconstruction method. In *2020 ieee international ultrasonics symposium (ius)* (pp. 1–4).

Byra, M., Sznajder, T., Korzinek, D., Piotrzkowska-Wróblewska, H., Dobruch-Sobczak, K., Nowicki, A., & Marasek, K. (2019). Impact of ultrasound image reconstruction method on breast lesion classification with deep learning. In *Iberian conference on pattern recognition and image analysis* (pp. 41–52).

Cheng, C.-T., Ho, T.-Y., Lee, T.-Y., Chang, C.-C., Chou, C.-C., Chen, C.-C., … Liao, C.-H. (2019). Application of a deep learning algorithm for detection and visualization of hip fractures on plain pelvic radiographs. *European radiology*, *29*(10), 5469–5477.

Cheng, J.-Z., Chou, Y.-H., Huang, C.-S., Chang, Y.-C., Tiu, C.-M., Chen, K.-W., & Chen, C.-M. (2010). Computer-aided us diagnosis of breast lesions by using cell-based contour grouping. *Radiology*, *255*(3), 746–754.

Cheng, J.-Z., Ni, D., Chou, Y.-H., Qin, J., Tiu, C.-M., Chang, Y.-C., … Chen, C.-M. (2016). Computer-aided diagnosis with deep learning architecture: applications to breast lesions in us images and pulmonary nodules in ct scans. *Scientific reports*, *6*(1), 1–13.

Collobert, R., & Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on machine learning* (pp. 160–167).

Commons, W. (2021). *File:aparelhodeultrassom.jpg — wikimedia commons, the free media repository.* Retrieved from https://commons.wikimedia.org/w/index.php?title=File:Aparelhodeultrassom.jpg&oldid=549084045 ([Online; accessed 19-November-2021])

Cruz, J. A., & Wishart, D. S. (2006). Applications of machine learning in cancer prediction and prognosis. *Cancer informatics*, *2*, 117693510600200030.

Czajkowska, J., Badura, P., Korzekwa, S., Płatkowska-Szczerek, A., & Słowińska, M. (2021). Deep learning-based high-frequency ultrasound skin image classification with multicriteria model evaluation. *Sensors*, *21*(17), 5846.

Das, A., & Rad, P. (2020). Opportunities and challenges in explainable artificial intelligence (xai): A survey. *arXiv preprint arXiv:2006.11371*.

Dastider, A. G., Sadik, F., & Fattah, S. A. (2021). An integrated autoencoder-based hybrid cnn-lstm model for covid-19 severity prediction from lung ultrasound. *Computers in Biology and Medicine*, *132*, 104296.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 ieee conference on computer vision and pattern recognition* (pp. 248–255).

Deng, L., Hinton, G., & Kingsbury, B. (2013). New types of deep neural network learning for speech recognition and related applications: An overview. In *2013 ieee international conference on acoustics, speech and signal processing* (pp. 8599–8603).

Dosovitskiy, A., & Brox, T. (2016). Inverting visual representations with convolutional networks. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 4829–4837).

Drukker, K., Sennett, C. A., & Giger, M. L. (2008). Automated method for improving system performance of computer-aided diagnosis in breast ultrasound. *IEEE Transactions on medical imaging*, *28*(1), 122–128.

Eskandari, A., Du, H., & AlZoubi, A. (2021). Towards linking cnn decisions with cancer signs for breast lesion classification from ultrasound images. In *Annual conference on medical image understanding and analysis* (pp. 423–437).

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *nature*, *542*(7639), 115-118.

Finlayson, S. G., Chung, H. W., Kohane, I. S., & Beam, A. L. (2018). Adversarial attacks against medical deep learning systems. *arXiv preprint arXiv:1804.05296*.

Giger, M. L., Karssemeijer, N., & Schnabel, J. A. (2013). Breast image analysis for risk assessment, detection, diagnosis, and treatment of cancer. *Annual review of biomedical engineering*, *15*, 327–357.

Girshick, R. (2015). Fast r-cnn. In *Proceedings of the ieee international conference on computer vision* (pp. 1440–1448).

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., … Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, *27*.

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Goudarzi, S., Asif, A., & Rivaz, H. (2020). Fast multi-focus ultrasound image recovery using generative adversarial networks. *IEEE Transactions on Computational Imaging*, *6*, 1272-1284. doi: 10.1109/TCI.2020.3019137

Gulli, A., & Pal, S. (2017). *Deep learning with keras*. Packt Publishing Ltd.

Habib, G., Kiryati, N., Sklair-Levy, M., Shalmon, A., Neiman, O. H., Weidenfeld, R. F., … Mayer, A. (2020). Automatic breast lesion classification by joint neural analysis of mammography and ultrasound. In *Multimodal learning for clinical decision support and clinical image-based procedures* (pp. 125–135). Springer.

Hall, T. J. (2003). Aapm/rsna physics tutorial for residents: topics in us: beyond the basics: elasticity imaging with us. *Radiographics*, *23*(6), 1657–1671.

Han, S., Kang, H.-K., Jeong, J.-Y., Park, M.-H., Kim, W., Bang, W.-C., & Seong, Y.-K. (2017a). A deep learning framework for supporting the classification of breast lesions in ultrasound images. *Physics in Medicine & Biology*, *62*(19), 7714.

Han, S., Kang, H.-K., Jeong, J.-Y., Park, M.-H., Kim, W., Bang, W.-C., & Seong, Y.-K. (2017b). A deep learning framework for supporting the classification of breast lesions in ultrasound images. *Physics in Medicine & Biology*, *62*(19), 7714.

He, K., Gkioxari, G., Dollár, P., Girshick, R., & R-CNN, M. (2017). 2017 ieee international conference on computer vision (iccv). *IEEE, Venice, Italy*, 2980–2988.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In

Proceedings of the ieee conference on computer vision and pattern recognition (pp. 770–778).

He, T., Guo, J., Chen, N., Xu, X., Wang, Z., Fu, K., … Yi, Z. (2019). Medimlp: using grad-cam to extract crucial variables for lung cancer postoperative complication prediction. *IEEE journal of biomedical and health informatics*, *24*(6), 1762–1771.

Heiss, G., Sharrett, A. R., Barnes, R., Chambless, L., Szklo, M., Alzola, C., & Investigators, A. (1991). Carotid atherosclerosis measured by b-mode ultrasound in populations: associations with cardiovascular risk factors in the aric study. *American journal of epidemiology*, *134*(3), 250–256.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, *30*.

Jalalian, A., Mashohor, S. B., Mahmud, H. R., Saripan, M. I. B., Ramli, A. R. B., & Karasfi, B. (2013). Computer-aided detection/diagnosis of breast cancer in mammography and ultrasound: a review. *Clinical imaging*, *37*(3), 420–426.

Jarosik, P., Klimonda, Z., Lewandowski, M., & Byra, M. (2020). Breast lesion classification based on ultrasonic radio-frequency signals using convolutional neural networks. *Biocybernetics and Biomedical Engineering*.

Jensen, J. A. (1991). A model for the propagation and scattering of ultrasound in tissue. *The Journal of the Acoustical Society of America*, *89*(1), 182–190.

Jin, C., Chen, W., Cao, Y., Xu, Z., Tan, Z., Zhang, X., … others (2020). Development and evaluation of an artificial intelligence system for covid-19 diagnosis. *Nature communications*, *11*(1), 1–14.

Kaproth-Joslin, K. A., Nicola, R., & Dogra, V. S. (2015). The history of us: from bats and boats to the bedside and beyond: Rsna centennial article. *Radiographics*, *35*(3), 960–970.

Ke, R., Bugeau, A., Papadakis, N., Kirkland, M., Schuetz, P., & Schönlieb, C.-B. (2021). Multi-task deep learning for image segmentation using recursive approximation tasks. *IEEE Transactions on Image Processing*, *30*, 3555–3567.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint* *arXiv:1412.6980*.

Koh, P. W., & Liang, P. (2017). Understanding black-box predictions via influence functions. In *International conference on machine learning* (pp. 1885–1894).

Kornecki, A. (2011). Current status of breast ultrasound. *Canadian Association of Radiologists Journal*, *62*(1), 31–40.

Krishnan, M. M. R., Banerjee, S., Chakraborty, C., Chakraborty, C., & Ray, A. K. (2010). Statistical analysis of mammographic features and its classification using support vector machine. *Expert Systems with Applications*, *37*(1), 470–478.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, *25*, 1097–1105.

Kurakin, A., Goodfellow, I., Bengio, S., & others. (2016). *Adversarial examples in the physical world.*

Lazo, J. F., Moccia, S., Frontoni, E., & De Momi, E. (2020). Comparison of different cnns for breast tumor classification from ultrasound images. *arXiv preprint arXiv:2012.14517*.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436–444.

Lin, Z., Li, S., Ni, D., Liao, Y., Wen, H., Du, J., ... Lei, B. (2019). Multi-task learning for quality assessment of fetal head ultrasound images. *Medical image analysis*, *58*, 101548.

Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, *16*(3), 31–57.

Liu, S., Johns, E., & Davison, A. J. (2019). End-to-end multi-task learning with attention. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 1871–1880).

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the ieee conference on computer vision and pattern recognition* (p. 3431-3440).

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Proceedings of the 31st international conference on neural information processing systems* (pp. 4768–4777).

Metzen, J. H., Genewein, T., Fischer, V., & Bischoff, B. (2017). On detecting adversarial perturbations. *arXiv preprint arXiv:1702.04267*.

Milletari, F., Navab, N., & Ahmadi, S.-A. (2016). V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3d vision (3dv)* (pp. 565–571).

Misra, S., Jeon, S., Managuli, R., Lee, S., Kim, G., Lee, S., … Kim, C. (2021). Ensemble transfer learning of elastography and b-mode breast ultrasound images. *arXiv preprint arXiv:2102.08567*.

Moon, W. K., Lee, Y.-W., Ke, H.-H., Lee, S. H., Huang, C.-S., & Chang, R.-F. (2020). Computer-aided diagnosis of breast ultrasound images using ensemble learning from convolutional neural networks. *Computer methods and programs in biomedicine*, *190*, 105361.

Moosavi-Dezfooli, S.-M., Fawzi, A., & Frossard, P. (2016). Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 2574–2582).

Nair, A. A., Tran, T. D., Reiter, A., & Bell, M. A. L. (2018). A deep learning based alternative to beamforming ultrasound images. In *2018 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 3359–3363).

Namburete, A. I., Xie, W., Yaqub, M., Zisserman, A., & Noble, J. A. (2018). Fully-automated alignment of 3d fetal brain ultrasound to a canonical reference space using multi-task learning. *Medical image analysis*, *46*, 1–14.

Narouze, S. N. (2018). *Atlas of ultrasound-guided procedures in interventional pain management*. Springer.

Nguyen, H.-G., Pica, A., Hrbacek, J., Weber, D. C., La Rosa, F., Schalenbourg, A., … Cuadra, M. B. (2019). A novel segmentation framework for uveal melanoma in magnetic resonance imaging based on class activation maps. In *International conference on medical imaging with deep learning* (pp. 370–379).

Noble, J. A., & Boukerroui, D. (2006). Ultrasound image segmentation: a survey. *IEEE Transactions on medical imaging*, *25*(8), 987–1010.

Oh, Y., Park, S., & Ye, J. C. (2020). Deep learning covid-19 features on cxr using limited training data sets. *IEEE transactions on medical imaging*, *39*(8), 2688–2700.

Panwar, H., Gupta, P., Siddiqui, M. K., Morales-Menendez, R., Bhardwaj, P., & Singh, V. (2020). A deep learning and grad-cam based color visualization approach for fast detection of covid-19 cases using chest x-ray and ct-scan images. *Chaos, Solitons & Fractals*, *140*, 110190.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … others (2011). Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, *12*, 2825–2830.

Rajpal, S., Lakhyani, N., Singh, A. K., Kohli, R., & Kumar, N. (2021). Using handpicked features in conjunction with resnet-50 for improved detection of covid-19 from chest x-ray images. *Chaos, Solitons & Fractals*, *145*, 110749.

Rasaee, H., & Rivaz, H. (2021). Explainable ai and susceptibility to adversarial attacks: a case study in classification of breast ultrasound images. In *2021 ieee international ultrasonics symposium (ius)* (p. 1-4). doi: 10.1109/IUS52206.2021.9593490

Rodrigues, P. S. (2017). Breast ultrasound image. *Mendeley Data*, *1*.

Rodríguez-Salas, D., Seuret, M., Vesal, S., & Maier, A. (2021). Ultrasound breast lesion detection using extracted attention maps from a weakly supervised convolutional neural network. In *Bildverarbeitung für die medizin 2021* (pp. 282–287). Springer.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International conference on medical image computing and computer-assisted intervention* (pp. 234–241).

Rouhafzay, G., Li, Y., Guan, H., Shu, C., Goubran, R., & Xi, P. (2020). An integrated deep architecture for lesion detection in breast mri. In *International conference on pattern recognition and artificial intelligence* (pp. 646–659).

Ruder, S. (2017). An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*.

Sánchez Fernández, I., Yang, E., Calvachi, P., Amengual-Gual, M., Wu, J. Y., Krueger, D., … others (2020). Deep learning in rare disease. detection of tubers in tuberous sclerosis complex. *PloS one*, *15*(4), e0232376.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam:

Visual explanations from deep networks via gradient-based localization. In *Proceedings of the ieee international conference on computer vision* (pp. 618–626).

Shankar, P. M., Dumane, V., Reid, J. M., Genis, V., Forsberg, F., Piccoli, C. W., & Goldberg, B. B. (2001). Classification of ultrasonic b-mode images of breast masses using nakagami distribution. *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*, *48*(2), 569–580.

Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). Deep inside convolutional networks: Visualising image classification models and saliency maps. In *In workshop at international conference on learning representations.*

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Singh, V. K., Rashwan, H. A., Abdel-Nasser, M., Sarker, M., Kamal, M., Akram, F., … Puig, D. (2019). An efficient solution for breast tumor segmentation and classification in ultrasound images using deep adversarial learning. *arXiv preprint arXiv:1907.00887*.

Standley, T., Zamir, A., Chen, D., Guibas, L., Malik, J., & Savarese, S. (2020). Which tasks should be learned together in multi-task learning? In *International conference on machine learning* (pp. 9120–9132).

Sudre, C. H., Li, W., Vercauteren, T., Ourselin, S., & Cardoso, M. J. (2017). Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep learning in medical image analysis and multimodal learning for clinical decision support* (pp. 240–248). Springer.

Sun, T., Zhang, R., Wang, J., Li, X., & Guo, X. (2013). Computer-aided diagnosis for early-stage lung cancer based on longitudinal and balanced data. *PloS one*, *8*(5), e63559.

Sun, Y., Wang, X., & Tang, X. (2013). Deep convolutional network cascade for facial point detection. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 3476–3483).

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2013). Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

Ting, F. F., Tan, Y. J., & Sim, K. S. (2019). Convolutional neural network improvement for breast cancer classification. *Expert Systems with Applications*, *120*, 103-115.

Umair, M., Khan, M. S., Ahmed, F., Baothman, F., Alqahtani, F., Alian, M., & Ahmad, J. (2021). Detection of covid-19 using transfer learning and grad-cam visualization on indigenously collected x-ray dataset. *Sensors*, *21*(17), 5813.

Vajihi, Z., Rosado-Mendez, I. M., Hall, T. J., & Rivaz, H. (2018). Low variance estimation of backscatter quantitative ultrasound parameters using dynamic programming. *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*, *65*(11), 2042–2053.

Van Ginneken, B., Schaefer-Prokop, C. M., & Prokop, M. (2011). Computer-aided diagnosis: how to move from the laboratory to the clinic. *Radiology*, *261*(3), 719–732.

van Sloun, R. J., & Demi, L. (2019). Localizing b-lines in lung ultrasonography by weakly supervised deep learning, in-vivo results. *IEEE journal of biomedical and health informatics*, *24*(4), 957–964.

Vishrutha, V., & Ravishankar, M. (2015). Early detection and classification of breast cancer. In *Proceedings of the 3rd international conference on frontiers of intelligent computing: Theory and applications (ficta) 2014* (pp. 413–419).

Wang, J., Yang, X., Cai, H., Tan, W., Jin, C., & Li, L. (2016). Discrimination of breast cancer with microcalcifications on mammography by deep learning. *Scientific reports*, *6*(1), 1–9.

Wang, P., Patel, V. M., & Hacihaliloglu, I. (2018). Simultaneous segmentation and classification of bone surfaces from ultrasound using a multi-feature guided cnn. In *International conference on medical image computing and computer-assisted intervention* (pp. 134–142).

Way, T. W., Hadjiiski, L. M., Sahiner, B., Chan, H.-P., Cascade, P. N., Kazerooni, E. A., ... Zhou, C. (2006). Computer-aided diagnosis of pulmonary nodules on ct scans: Segmentation and classification using 3d active contours. *Medical physics*, *33*(7Part1), 2323–2337.

Xie, X., Shi, F., Niu, J., & Tang, X. (2018). Breast ultrasound image classification and segmentation using convolutional neural networks. In *Pacific rim conference on multimedia* (pp. 200–211).

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In *European conference on computer vision* (pp. 818–833).

Zhang, Q., Wu, Y. N., & Zhu, S.-C. (2018). Interpretable convolutional neural networks. In

Proceedings of the ieee conference on computer vision and pattern recognition (pp. 8827–8836).

Zhang, Z., Luo, P., Loy, C. C., & Tang, X. (2014). Facial landmark detection by deep multi-task learning. In *European conference on computer vision* (pp. 94–108).

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2016). Learning deep features for discriminative localization. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 2921–2929).

Zhou, Y., Chen, H., Li, Y., Liu, Q., Xu, X., Wang, S., ... Shen, D. (2021). Multi-task learning for segmentation and classification of tumors in 3d automated breast ultrasound images. *Medical Image Analysis*, *70*, 101918.