Detection of Breast Cancer in Pocket Ultrasound Images Using Deep Learning

Freja Sahlin
Fr7123sa-s@student.lu.se

LUND UNIVERSITY

May 31, 2022

Master's thesis work carried out at
the Department of Mathematics, Lund University.

Supervisors: Ida Arvidsson, Ida.Arvidsson@math.lth.se
Kristina Lång, Kristina.Lang@med.lu.se
Jennie Karlsson, Jennie.Karlsson@math.lth.se

Examiner: Anders Heyden, Anders.Heyden@math.lth.se

Abstract

Breast cancer is the most frequently diagnosed form of cancer worldwide. 2 260 000 people were diagnosed with breast cancer year 2020, and 685 000 people deceased from it. In low income countries, breast cancer is commonly detected at a later stage when it is harder to treat, thus entailing a higher mortality rate. This is primarily due to the lack of knowledge and diagnostic tools available. A low cost breast diagnostic tool could therefore be a valuable solution in low income countries.

The objective of this thesis is to create a deep learning algorithm that can classify breast pocket ultrasound images as malignant, benign or normal. In this thesis, two different data sets were used. One ultrasound data set with 2062 images and one pocket ultrasound data set with 598 images. Four different approaches using convolution neural networks (CNN) were tested in order to produce the best model on the pocket ultrasound data set. In the first part, multiple different CNNs were created and trained on the ultrasound data set. The two models showing best results on the pocket ultrasound validation set were chosen for further evaluation. The second part consisted of augmentation of the ultrasound data set. The augmented images were then used to train the two chosen CNNs. In the third part, transfer learning was used in order to train the CNNs on both data sets. The last part of the thesis consisted of training the CNNs on the pocket ultrasound data set solely.

The best CNN gave an accuracy of 86.8% and an AUC value of 0.93 on the pocket ultrasound test set. This was achieved by training on both ultrasound and pocket ultrasound breast images using transfer learning. The performance on the pocket ultrasound test set was not improved by training the CNN's on augmented ultrasound images but training solely on pocket ultrasound images could be a good strategy with more data available. The results seem promising for the future and a perfected model, trained on more pocket ultrasound data, could possibly be implemented as a low cost diagnostic tool in countries without breast diagnostics.

Acknowledgements

I would like to express my deepest appreciation to my supervisor Ida Arvidsson for your guidance and valuable inputs in this thesis. Your support and creative ideas has been essential in the development of this thesis and it has given me great knowledge in the subject. I would also like to extend my appreciation to my clinical supervisor Kristina Lång for your encouragement and help with collecting the ultrasound and pocket ultrasound data. Our many interesting conversations has given me the insight and motivation to make this thesis as good as possible. I would also like to thank my assistant supervisor Jennie Karlsson for giving me good advice and knowledge on the subject and for your help with collecting the data sets.

Contents

Chapter 1
Introduction

The subject of this thesis is to create an automatic diagnostic tool for breast ultrasound images, specifically using *deep learning (DL)* to classify *pocket ultrasound* images as *malignant*, *benign* or normal. Health care providers today collects large amounts of data at a rate that traditional classifying methods cannot meet. This, in combination with the high performance of deep learning algorithms in image classification, makes the potential gain in the health care sector large [42].

1.1 Thesis Overview

This thesis is divided into four parts. The first part consists of an introduction to the subject, background information needed to understand the thesis and the data used (Chapters 1-3). In the second part, the methods used in the thesis is conveyed (Chapter 4). The third part (Chapter 5) presents the results produced and finally the fourth part (Chapter 6-7) discusses the results and potential future improvements as well as concluding the thesis.

Chapter 1 An overview of breast cancer, the relevance of the thesis and objective of the thesis is given.

Chapter 2 Background information about breast cancer and some basic concepts of deep learning is presented.

Chapter 3 A description of the different data sets used in the thesis.

Chapter 4 Presentation of the methodology of the thesis.

Chapter 5 The results of the thesis is presented.

Chapter 6 A discussion of the results of the thesis and potential future improvements is given.

Chapter 7 The conclusion, which can be drawn from the presented results, is stated.

1.2 Background

Breast cancer is the most common cancer type in the world. 2 260 000 people were diagnosed with breast cancer and 685 000 people deceased from it year 2020 [36]. In low income countries, breast cancer is commonly detected at a later stage thus entailing a higher mortality rate. This is primarily the result from lack of breast cancer awareness and limited resources for early detection and treatment [16]. Today, many women in low income countries avoid seeking medical care when finding a lump in the breast, due to fear of abandonment and a belief that cancer is an untreatable disease [14]. For women who seek medical care, the course of action is in almost all cases mastectomy due to lack of resources [48]. However, not all lumps are cancerous, many are benign *cysts* or *fibroadenoma*. Mastectomy is therefore an unethical and unsafe method to use without having a confirmed diagnosis and it can lead to other problems such as infection [30] [3]. *Mammography screening* has been shown to lower breast cancer specific mortality but the cost of a digital mammography machine is 20 000 to 50 000 euros [26] [16]. A mammography screening also needs a developed health care infrastructure. Therefore, it does not seem feasible to implement mammography in low resource settings.

A potential accessible and cost-effective tool to provide breast diagnostics in low resource settings is a *pocket ultrasound* probe, which costs around 4 000 euros [7]. Women that seek medical care for breast symptoms can then undergo diagnostic imaging at a lower cost. This would prevent unnecessary surgery for benign lesions whilst also detecting breast cancer in an early stage. By adding decision support to the pocket ultrasound using artificial intelligence, the tool can be more widely used, also by users with limited experience.

1.2.1 Purpose and Objective Statement

The purpose of this project is to evaluate if the pocket ultrasound probe Vscan Air from GE [8] in combination with a created deep learning classification algorithm could be a possible diagnostic tool in low income countries. The objective of the thesis is to produce a deep learning algorithm to classify pocket ultrasound images as malignant, benign or normal. In order to obtain this, the following problem statement will be analyzed and answered.

1.2.2 Problem Statement

- Is it possible to use ultrasound images as training data in order to classify pocket ultrasound images as malignant, benign or normal?
- Can the results on the pocket ultrasound images be improved by training on artificially augmented ultrasound images?
- Can the results on the pocket ultrasound images be improved by using transfer learning to train the created neural networks on both ultrasound and pocket ultrasound data?
- Is it possible to solely use pocket ultrasound images during training in order to classify pocket ultrasound images as malignant, benign or normal?
- Which of the created deep learning models gives the highest accuracy on the pocket ultrasound test set?

Chapter 2
Technical Background

2.1 Breast Cancer Screening

Breast cancer is the most common cancer type and the most common cause of cancer-related deaths for women globally [36]. The incidence is increasing due to, among other factors, an ageing population and a change in reproductive patterns [41]. Therefore, breast cancer has become a world wide health issue and the best way to decrease the mortality rate of breast cancer is to find them at an early stage when they are easier to treat. The most effective way to find the cancers at an early stage is by screening for breast cancer [10] [16].

2.1.1 Mammography

The most effective screening method to detect breast cancers at an early stage is with mammography screening. Mammography, an X-ray based method, can be used to detect cancer that has not yet started to cause symptoms. Mammography screening can therefore reduce the mortality of breast cancer through early detection [16]. However, many cancers are missed in mammography, especially cancers in dense breasts where the *sensitivity* can be as low as 30-48% [4]. This is due to overlaying dense breast tissue that makes cancers harder to visualize in mammography images. Women with dense breast tissue also have a 4-6 times higher risk of getting breast cancer compared with women with

entirely fatty breasts and the prevalence of dense breasts is about 40% of women in ages 40-74 in Sweden[5].

2.1.2 Ultrasound

An attractive supplement to mammography, especially for patients with dense breast tissue, is ultrasound since it visualizes the tissue without overlap and is more affordable and tolerable by patients [20]. An ultrasound image is created by high frequency sound waves that are transmitted by the ultrasound probe. The sound waves will penetrate the tissue until a mass or object appears. Some of the sound waves will then be reflected back and the reflected sound waves are received by the ultrasound probe. A connected computer then constructs an image out of the time of each echos return and the speed of sound in the tissue. A traditional ultrasound machine consists of a probe with a cord to a computer, where the computations are made, and a screen where the image is shown in real time [12]. The resolution of an ultrasound image is divided into lateral and axial resolution. The axial resolution depends on the frequency of the sound waves, a high frequency entails higher attenuation in soft tissue which leads to poor axial resolution in deeper tissue. If an object is located further down in the tissue, a lower frequency needs to be used at the expense of axial resolution. Lateral resolution depends on the focus of the sound beam, a smaller width of the beam entails a higher lateral resolution whilst a wider width gives a larger view of the tissue [46].

2.1.3 Pocket Ultrasound

The pocket ultrasound device used in this thesis is the Vscan Air ultrasound probe developed by GE. The device is a wireless ultrasound probe that connects to a smartphone with bluetooth where the ultrasound image is shown in real time. The name 'pocket ultrasound' comes from the fact that it is an ultrasound system that can fit in your pocket. The probe has two different sides for different purposes. One side is a curved ultrasound probe that sends lower frequency sound waves and the other side is a linear ultrasound probe with higher frequency. The technique of creating the pocket ultrasound images is based on the same principles as in an ordinary ultrasound system [8]. Figure 2.1 displays the Vscan Air probe from GE with measurements.

2.1.4 Tumour Characteristics

When finding an object in the breast, the doctors want to determine if the finding is malignant or benign. Malignant means cancerous and benign means a non-cancerous finding. When determining if a finding is malignant or benign, characteristics such as

Figure 2.1: Picture and measurements of the Vscan Air ultrasound probe [8]. The illustration is used with permission from GE.

shape, *acoustic enhancement/shadowing*, *echogenity* and *margins* are evaluated [35].

Shape
When looking at the shape of a finding a malignant tumour usually have an irregular shape as can be seen in Figure 2.2 (b). A benign finding on the other hand usually has a regular shape, as displayed in Figure 2.2 (a).

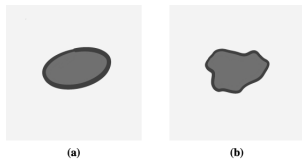

Figure 2.2: Image of typical benign shape (a) and malignant shape (b) [18].

Margin
The margin of a finding refers to if the object is circumscribed or not.
A malignant finding usually have an unclear margin without clear circumscription whilst a benign finding has a smooth margin. Therefore it is a good tool for determining if a tumour shown in an ultrasound image is malignant or not. This can be seen in Figure 2.3.

Acoustic Enhancement and Shadowing
Acoustic enhancement typically appears in a breast ultrasound image in water filled findings such as cysts. This is due to the low attenuation of sound in fluid which causes

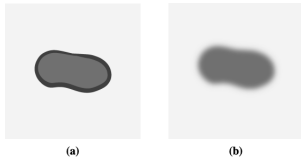

Figure 2.3: Image of circumscribed and uncircumscribed margins typical for benign (a) and malignant (b) [18].

the tissue below to appear lighter underneath the finding [31]. Acoustic shadowing is seen under solid mass findings. A solid mass will often have a high attenuation thus entailing a shadowing underneath the finding [44]. These phenomenons are shown in Figure 2.4.

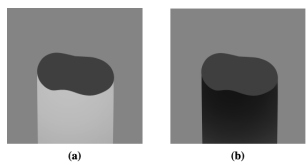

Figure 2.4: Image of aucoustic enhancement (a) and acoustic shadowing (b) of a finding [18].

Echoic Pattern

Echoic pattern is a good tool in breast ultrasound images to determine if a finding is malignant or not. An echoic finding, shown in Figure 2.5 (a), will appear dark grey in the ultrasound image and it is a common characteristic of a malignant tumour. A benign cyst on the other hand is anechoic and will appear black in an ultrasound image, see Figure 2.5 (b).

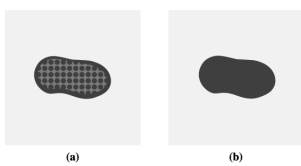

Figure 2.5: Image of an echoic (a) and anechoic (b) mass [18].

2.2 Deep Learning

AI is a broad term used in various ways today as well as its definition. AI is essentially a machine that performs a task in a "smart" way. Machine learning is a sub-field in AI where a machine teaches itself with help from designed features to perform a task by training on a set of data and then ideally performing the same on new data [15]. Furthermore, artificial neural networks (ANN) is a segment within machine learning. ANN is a set of algorithms inspired by the human brain and it consists of multiple nodes, *neurons*, that are interconnected similar to neurons in a human brain [19] [34]. An ANN with more then three layers can be considered a deep learning algorithm, a branch within ANN that refers to 'deeper' neural networks [19]. The existence of deep learning goes back to 1940 but has become essential in recent years due to massive collection of data and advanced computer power that allows networks to be larger and deeper. These factors are the reason for the high accuracy that deep learning algorithms can achieve today [15]. The relationship between AI, machine learning, artificial neural networks and deep learning is illustrated in Figure 2.6.

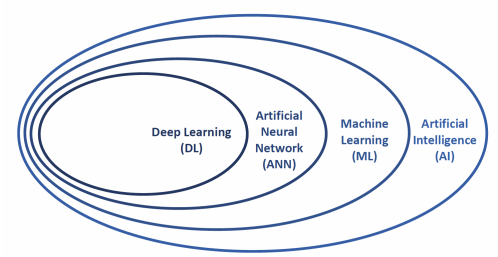

Figure 2.6: Relationship between artificial intelligence, machine learning, artificial neural networks and deep learning shown in euler graph.

2.2.1 Multi Layer Perceptron

A multi-layer perceptron (MLP) is a type of ANN with one or more fully connected hidden layers of type *feed forward*, meaning the input propagates forward in the layers to the output as Figure 2.7 displays. Before deep learning became popular the MLP networks usually consisted of two hidden layers but today many more layers are often used [34].
Figure 2.7 illustrates a MLP with three input nodes, two fully connected hidden layers and three output nodes. Since all layers are fully connected in a MLP, the computational problem becomes large with many nodes. For example in image classification,

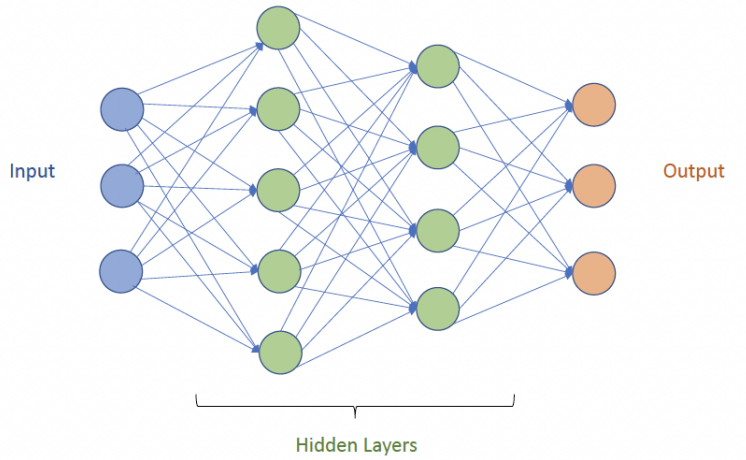

Figure 2.7: Illustration of a multi-layer perceptron with two hidden layers and three output classes.

all pixels in an image would be connected to all hidden nodes. Furthermore the MLP does not take *spatial information*, such as information from neighbouring inputs, to account. Therefore, the MLP is not the optimal method to use in image classification problems [11] [34]. Equation (2.1) shows the calculation of a MLP's output y in one layer from input x, activation function φ and weights w. n in the formula is the number of input nodes. Information about activation functions and weights is found in the section below.

$$y = \varphi(\sum_{i=0}^{n} w_i x_i) \tag{2.1}$$

2.2.2 Convolutional Neural Networks

Convolutional neural networks (CNN), are commonly used in image classification since they are good at recognizing spatial relations in inputs, for example information from neighbouring pixels in an image. The CNN focuses mainly on *feature detection* in multiple layers. The first layers detects small features such as brightness or lines that are passed to the next layer, which builds more complex features out of the ones passed from the layer before [34].

Activation Function

The activation function plays an important role in artificial neural networks and it can improve the performance noticeably if chosen correctly. Some common activation functions for CNN models are ReLU, tanh and sigmoid shown in Equations (2.3),(2.4) and (2.5) [17].

$$f(x) = \begin{cases} 0 & x < 0 \\ x & x \geq 0 \end{cases} \tag{2.2}$$

$$f(x) = \frac{2}{1 + e^{-2x}} - 1 \tag{2.3}$$

$$f(x) = \frac{1}{1 + e^{-x}} \tag{2.4}$$

Weights

Weights are numerical values that determines how important the input is to the output, a small weight entails less influence on the output and a large value entails a large influence on the output [29].

Kernel

A kernel is a matrix of weights that together with an activation function is a filter that extracts features from the layer before. In a CNN, the weights in the kernel is decided and updated by the model itself during training [34].

Convolutional Layer

The fundamental part of a CNN is the mathematical *convolution*. Equation (2.2) displays the convolutional computation H of a 2D image I with pixel coordinates (i,j) and a kernel K with dimensions (m,n) [34].

$$H(i,j) = (I * K)(i,j) = \sum_{m,n} I(i-m, j-n)K(m.n) \tag{2.5}$$

Pooling Layer

A pooling layer consists of a pooling filter without weights, that summarizes the features created from a convolutional layer. The function of a pooling layer is to reduce the dimensionality of the feature maps and thus also reduce the complexity of the computations as well as make the model invariant to the position of the features. *Max pooling* layers only store the highest feature value from a certain area in the feature map and it is often used in combination with convolutional layers [34]. Figure 2.8 shows a graphical explanation of the convolutional layer with an input image and a kernel with size 3 x 3, marked red in the image, as well as a max pooling layer.

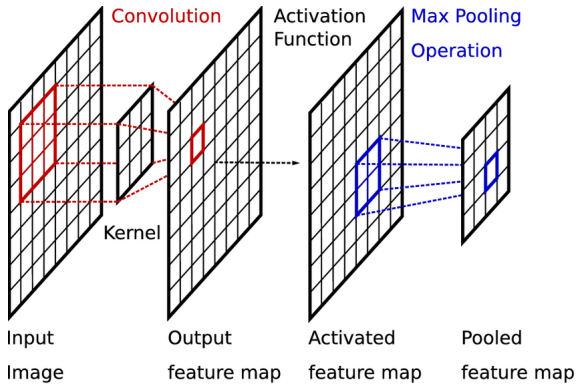

Figure 2.8: Visual representation of a convolutional layer with kernel 3 x 3 and a max pooling layer with pool size 2 x 2 [13]. The image is added to this thesis in consent with the terms in the Creative Commons license [28].

Dropout Layer

Dropout is a useful tool to use during training in order to reduce overfitting, which is when the model becomes to adapted to the training data. In the dropout layer some nodes are randomly dropped, which reduces the risk of a few nodes dominating the result of the output and therefore overfitting to the training data [49]. When designing a dropout layer, a probability is chosen which is the likelihood that a node will be dropped.

Loss Function

The loss function is a way to determine how good the model is. It measures the distance between the output of the model and the expected output. In multiple class problems the loss *categorical cross entropy* function (E) is used and the equation is shown in Equation (2.6). n is the number of classes in the data, in our case three: malignant, benign and normal. y is the true label and $\hat{y}$ is the probability of a sample belonging to class i [6].

$$E = -\sum_{i=0}^{n} y_i log(\hat{y}_i) \tag{2.6}$$

2.3 Training a CNN

When training a machine learning algorithm it is challenging to foresee how well the model performs on unseen data. Therefore, to evaluate this, the data set is divided into three parts, *training, validation* and *testing*. The training data set is given to the algorithm to train and learn from. The validation data set is used to estimate a generalized performance on data not seen in training and to change *hyper parameters* to make the model better. The performance on the validation data can however not surely be reproduced on new unseen data since the hyper parameters are optimized based on the results from the validation data. To be able to rely on the performance of a model it is tested on the test set which then gives the generalized performance on new unseen data that has not been used in any part of the training or optimization process [34]. The method of splitting the data set can be seen in Figure 2.9

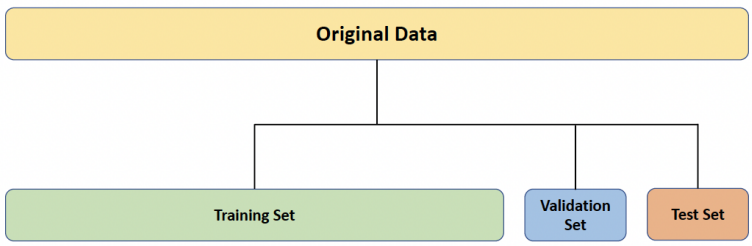

Figure 2.9: Visual representation of dividing data set into training, validation and test sets.

2.3.1 K-fold Cross Validation

K-fold cross validation is a method to get reliable validation results which is especially useful in cases of small data sets. This is due to the validation data set then being too small to rely on during the training process. In k-fold cross validation the data set is divided into k parts or 'folds' where one fold is used for validation and the others are used for training. K slightly different models will be produced due to k slightly different training and validation sets. In the first model, the first fold is used for validation, in the second model, the second fold is used as validation and so on until all folds have been used as validation to produce k models. The estimated generalized performance will then be the average of the k validation results [34]. Standard deviation and mean value can be calculated from the K models in order to evaluate how the performance differs. A graphical explanation of the k-fold cross validation method is shown in Figure 2.10.

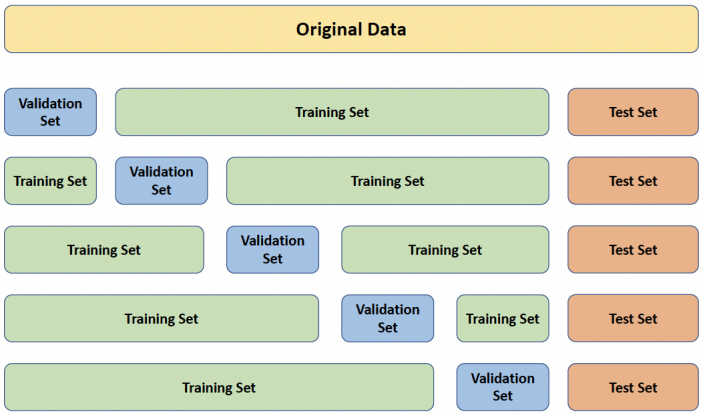

Figure 2.10: Visual representation of the k-fold cross validation method with k = 5 folds.

2.3.2 Transfer Learning

Transfer learning is a method where an already trained model or *pre-trained model*, is used as a starting point for another model which makes the learning process less difficult. In that way features of one problem can be reused for prediction in a new similar problem.

Typically the first layers in the pre-trained model is reused and the last layers are trained on the new data set, thus using the features from the pre-trained model and optimizing them to the new problem [43]. Some common transfer learning networks available in the Keras framework are VGG16 [24], InceptionV3 [23], InceptionResNetV2 [22] and Xception [25]. These networks are trained on the ImageNet data set, which is a data set commonly used in deep learning research and consists of more than 14 000 000 images of almost 22 000 different classes [45].

2.3.3 Data augmentation

Data augmentation is when small changes are made to the data which allows the algorithm to see it as a different data point. For example images can be augmented by adding a zoom, changing the intensity in the image or adding noise to the image. This makes the data set larger and can force the algorithm to look at important features instead of for example noise.

2.3.4 Hyper Parameters

Hyper parameters are changeable parameters that are altered in order to get the best results on the validation data. These differ from for example the training weights that are chosen and optimized by the model to perform good on the training data.

Batch Size and Epochs

Batch size is used in certain *optimizing techniques* and refers to how much data from the data set is used before the weights are updated during training. Instead of processing the whole data set at once, it is divided into *batches* of data used for training. After processing one batch, the weights are updated. When the whole data set has been used for training one *epoch* has been completed. The number of epochs used in the model usually depends on the *validation loss*, the loss function from the validation data set. The number of epochs to be run is chosen before the training process and *early stopping* can be implemented to stop the model when the validation loss is no longer improving [6].

Learning Rate and Optimization Techniques
The learning rate is a positive value, usually between 0 and 1, used in the *optimizing technique* to decide how much the weights should be updated per batch during training. There are different optimizing techniques when training a model and a common one is stochastic gradient decent. The method minimizes the loss function from the *gradient* of P examples, one batch, in the training data [34]. Equation (2.7) displays the weight update from one sample i in the batch. η is the learning rate and $\frac{\delta E}{\delta w_i}$ is the gradient of the loss function with weight w_i.

$$w_i \rightarrow w_i - \eta \frac{\delta E}{\delta w_i} \tag{2.7}$$

Another optimization technique is *Adaptive Moment Estimation* (ADAM), which have shown success in multiple areas including image classification with CNN. ADAM is similar to the stochastic gradient descent method but updates the learning rate of the model depending on the past gradients [51].

Class Weights
Class weights is a hyper parameter that allows all classes to be equally important during training. It is also possible to make one class more important then others if needed. Class weights is useful when the data set is imbalanced since the most frequent class naturally will influence the result more. To make all classes influence the result equally much the less frequent classes will be weighted up [47]. In this thesis, the classes will be weighted based on the frequency of their occurrence by using the argument 'balanced' the scikit-learn build in method 'compute_class_weights' [27].

2.3.5 Python

Python is a popular programming language to use when developing deep learning projects, one reason for it being the large amount of available *libraries* and *frameworks*. This allows the user to access many functions that makes it easier to solve the given problem. A commonly used framework in deep learning projects is Keras [21] and it is also used in this project. Some standard libraries in python that is commonly used in deep learning is

NumPy [32] for computations and Scikit-Learn [40] for data mining and analyzing [33].

2.4 Evaluation Metrics

2.4.1 Sensitivity, Specificity and Accuracy

Sensitivity and specificity are often used as validation or test performance measures. To calculate these metrics *true positives (TP), true negatives (TN), false positives (FP)* and *false negatives (FN)* are needed [34]. Sensitivity is defined as *the fraction of actual positives that were correctly predicted* and the definition is shown in Equation (2.8). Specificity is defined as *the fraction of actual negatives that were correctly predicted*, definition in Equation (2.9).

$$Sensitivity = \frac{TP}{TP + FN} \tag{2.8}$$

$$Specificity = \frac{TN}{TN + FP} \tag{2.9}$$

2.4.2 Confusion Matrix and Accuracy

A confusion matrix is another method for evaluating the performance of a model, for example the performance on the validation or test set. In binary classifications, the confusion matrix displays the number of TP, TN, FP and FN [34]. This can be seen in Figure 2.11 (a). Figure 2.11 (b) displays a confusion matrix of three classes. Accuracy is another common evaluation metric and it is defined as the *fraction of correct predictions* out of all predictions and the definition is shown in Equation (2.10) [34].

$$Accuracy = \frac{TN + TP}{TN + FP + TP + FN} \tag{2.10}$$

In order to calculate the accuracy from Figure 2.11 (b), the number of correct classifications are divided by the total amount of predictions, as can be seen in Equation (2.11). If the number of samples differ between classes, the *weighted accuracy* can be a better performance measurement.
The weighted accuracy is calculated by taking the sensitivity for each class divided by the total number of classes. Equation (2.12) displays the calculation of the weighted accuracy for the confusion matrix in Figure 2.11 (b).

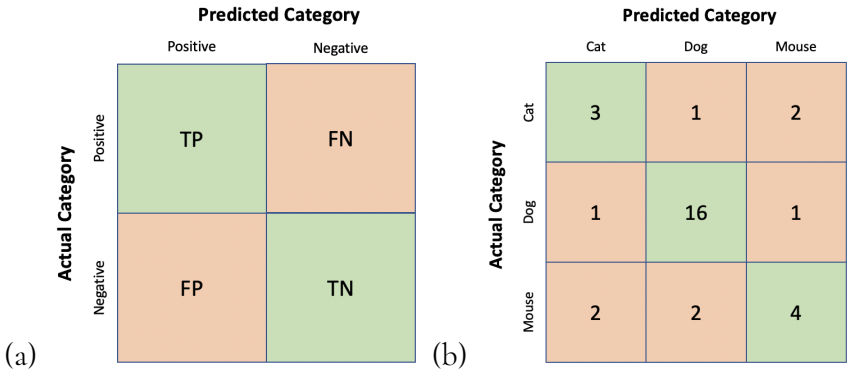

Figure 2.11: Confusion matrix with true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), for a binary classification problem.

$$Accuracy: \quad \frac{3+16+4}{6+18+8} = 0.72 \tag{2.11}$$

$$Weighted\ Accuracy: \quad \frac{0.5+0.89+0.5}{3} = 0.63 \tag{2.12}$$

2.4.3 ROC Curve and AUC

The *Receiver Operator Characteristic* (ROC) curve is another way to measure the performance in binary classification problems as well as the *Area Under the ROC Curve (AUC)*. The ROC curve is created by changing the *threshold* of predicted classes and plotting the sensitivity against 1-specificity for each change in threshold value [34]. Threshold is the value that decides which class a sample should belong to. The AUC value is the area under the plotted curve and a higher value implies a better classifying model. The AUC value is in the range 0 to 1 where a value of one implies a perfect classifier. Figure 2.12 illustrates three different ROC curves and their AUC values where the yellow curve represents the best classifier and the green represents random guessing.

In order to plot the ROC curve with three classes in this thesis, the malignant class was plotted against the benign and normal class. This was done in all parts of the project.

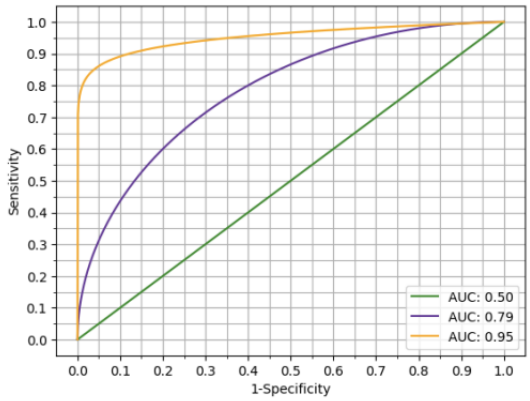

Figure 2.12: Image of three different ROC curves and their AUC values [2]

Chapter 3
Data

Two types of data sets were used in this thesis, one ultrasound data set and one pocket ultrasound data set. Both data sets consist of three classes, malignant, benign and normal images. A detailed description of the data sets will be given in the section below.

3.1 Ultrasound Data Set

Ultrasound data was collected retrospectively from three different sources: Swedish, Egyptian and Dutch data. The ultrasound data set consisted of three different ultrasound data sets, one Dutch, one Swedish and one data set from Egypt.

3.1.1 Swedish Data Set

The Swedish data set consisted of 481 images acquired from women in Malmö, Sweden year 2018 and retrieved during spring 2022. The study was approved by the Ethics Review Authority (2019-04607) and informed consent was waived for retrospectively collected ultrasound data. The ultrasound images were acquired with the ultrasound system Logiq E9 och Logiq E10 by Unilabs Mammography Unit at Skåne University Hospital. Table 3.1 displays the distribution of the different classes in the data set. Some patients from spring 2022 were also retrieved.

For these patients, pocket ultrasound images were also captured. Table 3.1 displays the distribution of the Swedish data set.

Table 3.1: Class distribution of the Swedish data set.

Class	Number of images
Malignant	164
Benign	121
Normal	196
Total	481

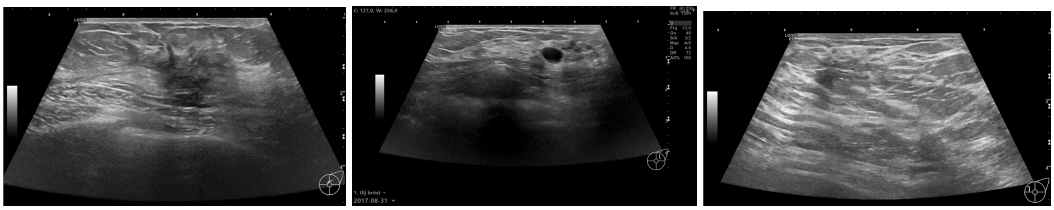

Figure 3.1: Example of a malignant, benign and normal image from the Swedish data set.

3.1.2 Egyptian Data Set

The Egyptian data set consists of 780 images taken from 600 women by Cairo University at Baheya Hospital in Egypt 2018 [1]. Table 3.2 displays the distribution of the Egyptian data set.

Table 3.2: Class distribution of the Egyptian data set.

Class	Number of images
Malignant	210
Benign	437
Normal	133
Total	780

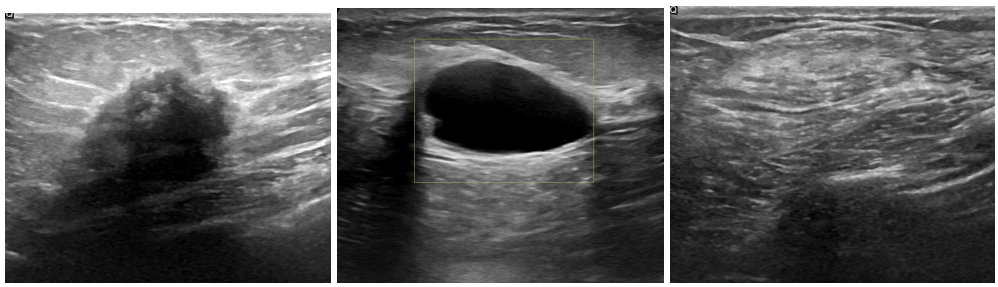

Figure 3.2: Example of a malignant, benign and normal image from the Egyptian data set.

3.1.3 Dutch Data Set

The Dutch data set consists of 801 breast ultrasound images collected from the website ultrasoundcases.info with permission from the owners. The website is a collaborative work between FUJIFILM Healthcare Europe and Sonoskills. Sonoskills is the leading ultrasound learning provider in Europe and FUJIFILM is a medical imaging company. The data set only contained benign and malignant images and the normal images were created by cropping normal parts of the benign and malignant images with consultation from radiologist Kristina Lång. Table 3.3 displays the distribution of the Dutch data set.

Table 3.3: Class distribution of the Dutch data set.

Class	Number of images
Malignant	388
Benign	312
Normal	101
Total	801

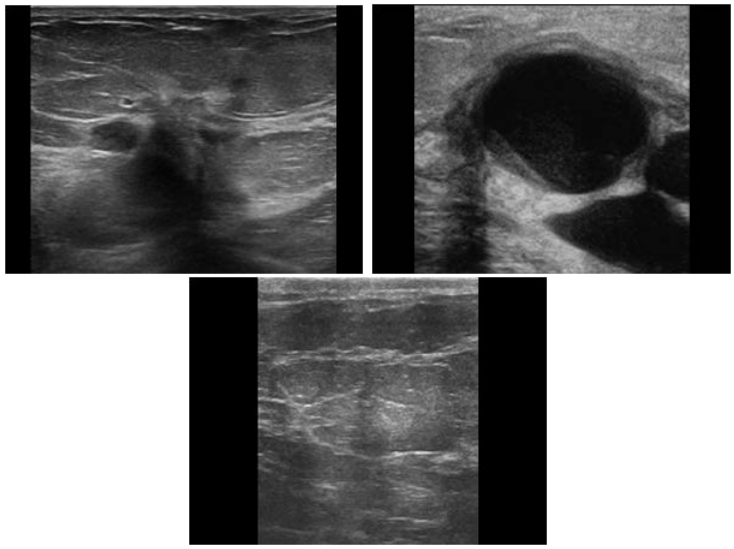

Figure 3.3: Example of a malignant, benign and normal image from the Dutch data set.

3.2 Pocket Ultrasound Data Set

The pocket ultrasound data set were collected prospectively during January-March 2022 at Unilabs Mammography Unit at Skåne University Hospital in Sweden. The stud in- form were take ages are the pock ltra- sound data set collected.

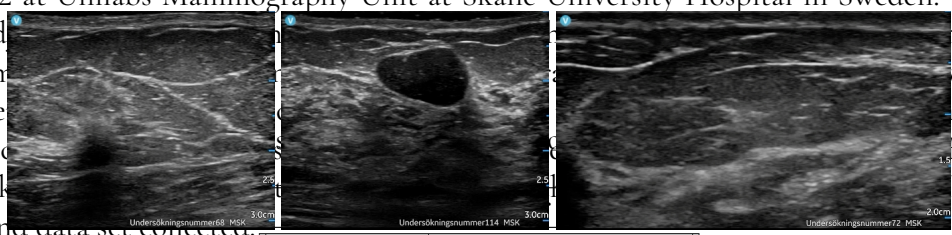

Figure 3.4: Example of a malignant, benign and normal image from the pocket ultrasound data set.

Malignant	91
Benign	167
Normal	340
Total	598

Chapter 4
Methods

The project was divided into four different parts. The first part consisted of creating CNN's and training them on ultrasound data only. The two best performing models on the pocket validation set were chosen for further evaluation and were used in all parts of the project. In the second part, differences in the pocket ultrasound and ultrasound images were examined and the two CNN's from part one were trained on augmented ultrasound images. The third part consisted of partly training on ultrasound images and partly training on pocket ultrasound images using transfer learning. In the last part of the project, the CNN's were trained on pocket ultrasound images only. In all parts of the project k-fold cross validation was used and the models were validated and tested on ultrasound and pocket ultrasound images.

4.1 Pre-Processing Data

4.1.1 Data Set

The three ultrasound data sets were added together to one with the specifications presented in Table 4.1. In this thesis I will refer to the joined ultrasound data set as 'the ultrasound data set'.

4.1.2 Zero Padding Images

The ultrasound and pocket ultrasound images are rectangular and differ in size. Since the neural networks used takes quadratic images, the size of the ultrasound and pocket

Table 4.1: Class distribution of joined ultrasound data set.

Class	Number of images
Malignant	762
Benign	870
Normal	430
Total	2062

ultrasound images were modified. In order to resize the images from rectangular to quadratic form without effecting the biological properties in the images, zero padding was used. Figure 4.1 illustrates an image before and after zero padding.

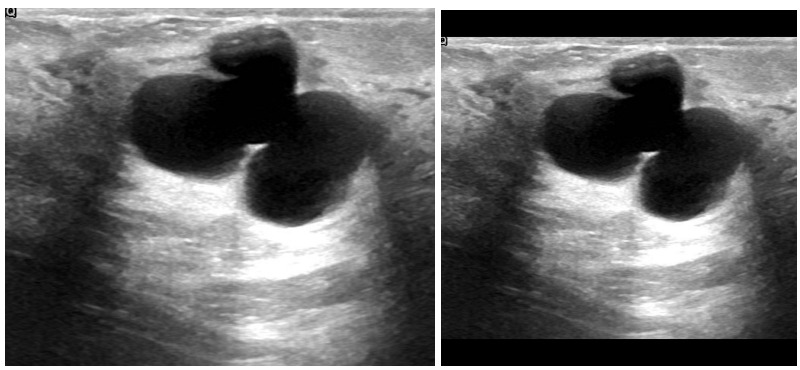

Figure 4.1: Image before and after zero padding.

4.1.3 K-fold Cross Validation and Test Sets

Two test sets were set aside, one from the joined ultrasound data set and one from the pocket ultrasound data set. The ultrasound test set consisted of 10% of the ultrasound data set, 206 images, and the pocket ultrasound test set consisted of around 25% of the pocket ultrasound data set, 152 images. After setting aside the test sets, the ultrasound and pocket ultrasound data set was divided into five parts using the k-fold cross validation method. In all parts of the project the ultrasound data set folds were tied to the pocket ultrasound data set folds, for example when training model 1, the ultrasound validation set comes from joined fold 1, the ultrasound training set comes from what is left after setting aside fold 1 and the pocket ultrasound validation set comes from fold 1 in the pocket ultrasound set of images. A graphical explanation for this can be seen in Figure 4.2.

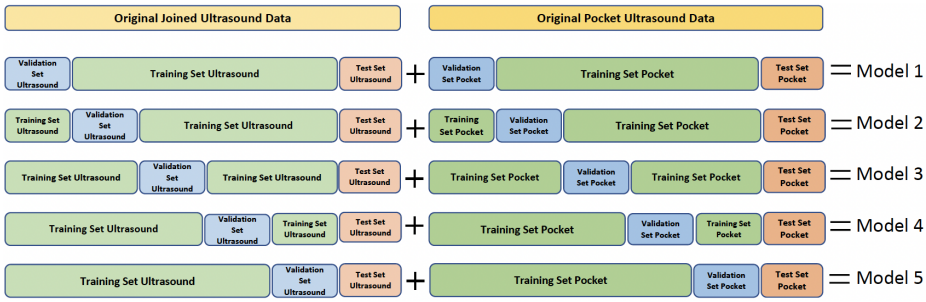

Figure 4.2: Illustration of the training process with 5-fold cross validation on data sets joined and pocket. One model consists of a training set of either ultrasound images or pocket ultrasound images or both, two validation and test sets, one ultrasound and one pocket.

4.2 Part 1: Training on Ultrasound Images

In this part of the project the models were trained on zeropadded ultrasound images. Multiple different neural networks were created and tested, among them were transfer networks InceptionV3, InceptionResNetV2, Xception and VGG16. The two models with most potential on the pocket ultrasound validation set was selected for further optimization. The architecture of these models is described and illustrated in the sections below. During training, the pocket ultrasound validation set was used to determine hyper parameters and optimal epoch to stop the training. Five models were created for each type of network, one for each fold in 5-fold cross validation. The average pocket ultrasound and ultrasound validation accuracy for the five models was calculated as well as the standard deviation of the pocket ultrasound validation set and the ultrasound training accuracy. The models were then tested on the two test sets. From testing the average pocket ultrasound and ultrasound accuracies for the five models were calculated as well as the weighted pocket ultrasound accuracy, standard deviation and AUC value. The confusion matrix and ROC curve was plotted from the pocket ultrasound test predictions.

4.2.1 Architecture of the Simple CNN

In this report I will refer to the simple convolutional neural network used as 'the simple CNN'. The hyper parameters used in this part was the following: input size 180x180x3, activation function was ReLU on hidden layers and softmax on output layer. Kernel size 3x3 in convolutional layers and and pool size 2x2 in max-pooling layers. Dropout layer had 0.5 probability, batch size was 32 and early stopping was used. Figure 4.3

displays the architecture of the simple CNN.

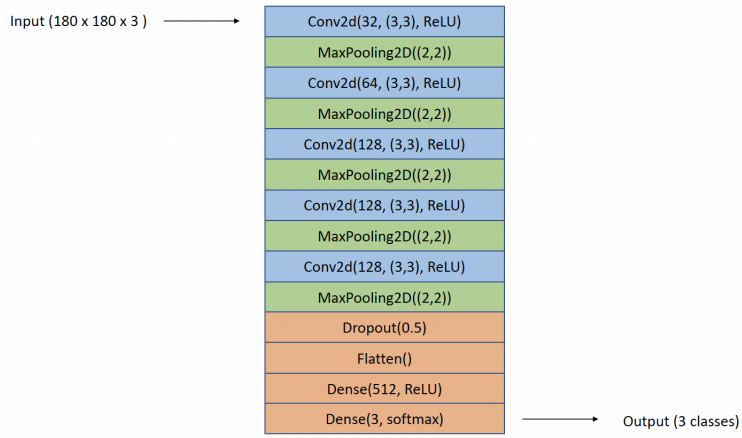

Figure 4.3: Illustration of the simple CNNs' architecture.

4.2.2 Architecture of VGG16

The neural network trained after the VGG16 network had the following hyper parameters: Input size 5x5x512, same as the output from VGG16, activation function was ReLU on hidden layers and softmax on output layer, kernel size was 3x3 and 2x2, batch size was 32 and early stopping was used. Figure 4.4 displays the architecture of the network trained after the VGG16 network.

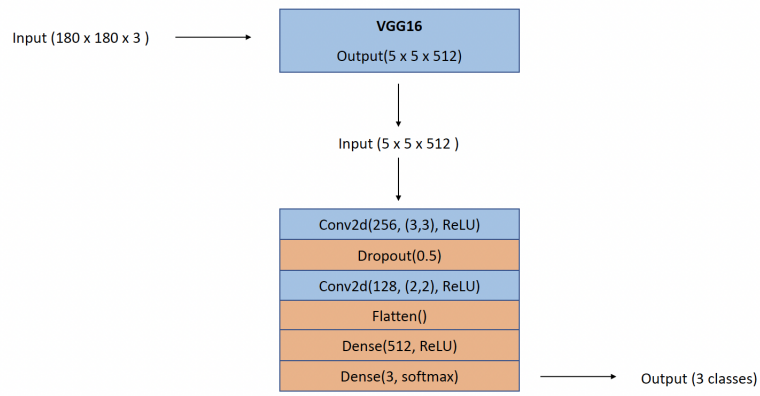

Figure 4.4: Illustration of the VGG16 architecture.

4.3 Part 2: Analyzing Properties in Ultrasound and Pocket Ultrasound Images

4.3.1 Histogram

In order to analyze the intensity difference between the two types of images, intensity histograms were made. Two images, one ultrasound and one pocket ultrasound image, were collected from the same patient on the same area for each class malignant, benign and normal. Two normalized pixel intensity histograms, one for ultrasound and one for pocket ultrasound, were made for each class and compared. Another two average pixel intensity histograms were made over all ultrasound and pocket ultrasound images and the result was compared to the single class and image histograms.

4.3.2 Intensity and Noise Augmentation

Before augmenting the data set, k-fold cross validation was done and the validation sets for each fold was not included in the augmentation process since the validation sets should resemble the test sets. The intensity in the five training ultrasound folds were then brightened and darkened using the 'ImageEnhance.Brightness' method from the 'PIL' library [37]. The pixel intensity was changed to 0.8 times the original value and 1.2 times the original value. The augmented images was added together with the original ultrasound images to a new data set, keeping the 5 folds separate. Two types of noise was also added to the ultrasound folds, gaussian noise and local variance (localvar) noise by using the 'random_noise' method in the 'skimage' library [39]. Local variance noise is gaussian distributed noise with a variance dependant on the intensity of the image [39]. The images with gaussian and local variance noise were added to the new data set containing the original ultrasound images, brighter and darker ultrasound images and gaussian and local variance noise ultrasound images. The new data set was thereby five times the size of the original ultrasound data set with five training folds, zero padding was added after the augmentation of the images to avoid changing the black pixels outside of the image. Figure 4.5 displays an example of an ultrasound image before and after adding gaussian and localvar noise.

4.3.3 Training on Augmented Ultrasound Images

The two models described in section 4.2.1 and 4.2.2 were trained on zeropadded and augmented ultrasound images. The unaugmented pocket ultrasound validation set

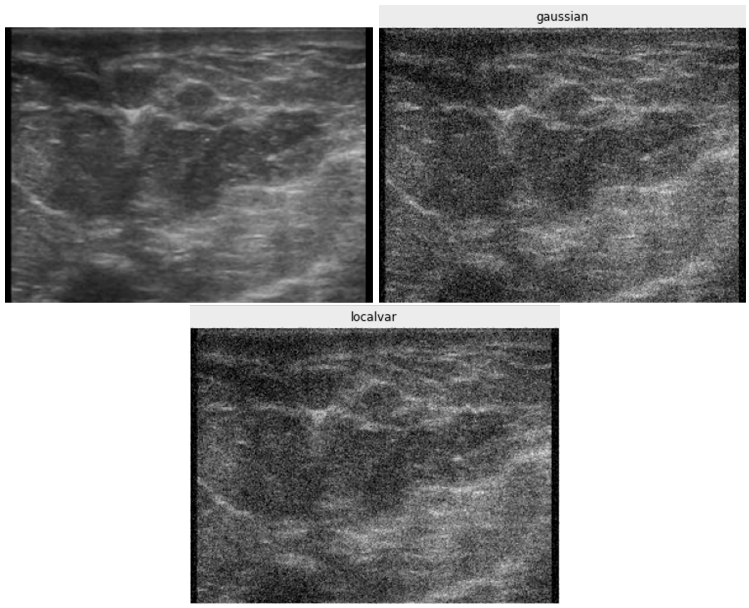

Figure 4.5: Illustration of how an image can look with gaussian and localvar noise.

was used to optimize the models and the same evaluation metrics as in part one was produced for validation and testing and the results were compared.

4.4 Part 3: Training using Transfer Learning

In the third part, the ultrasound images were used as well as the pocket ultrasound images during training using transfer learning. The models from part one were used and the last layers of the models were re-trained on the pocket ultrasound images. Figure 4.7 displays a graphical explanation of the transfer learning in part three. The evaluation metrics produced in part one and two were produced after validation and testing and the results were compared.

4.5 Part 4: Training on Pocket Ultrasound Images

In the last part of the project the two types of networks were trained solely on pocket ultrasound images and then validated and tested on both pocket ultrasound and ul-

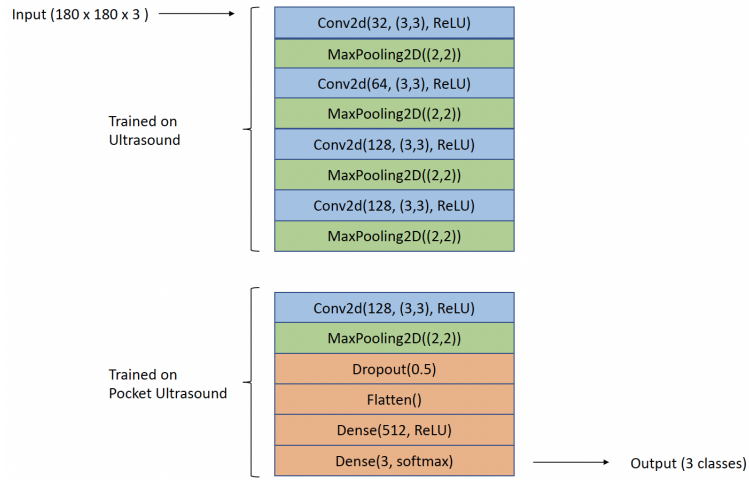

Figure 4.6: Illustration of transfer learning in the simple CNN.

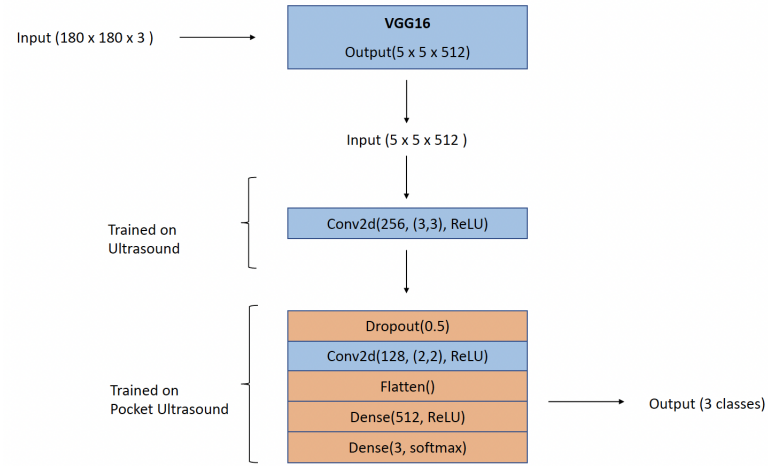

Figure 4.7: Illustration of transfer learning in the VGG16 network.

trasound images. The same evaluation metrics as in the other parts, were produced.

Chapter 5
Results

5.1 Part 1: Training on Ultrasound Images

The simple CNN and the transfer network VGG16 was used in all 4 parts of the project. In the first part of the project, where the networks were trained on ultrasound images, the VGG16 got the highest accuracy on the validation pocket data set, which is displayed in Table 5.1. It also has the lowest standard deviation, meaning the pocket validation accuracy of the five models in VGG16 differ less from model to model then the five simple CNN models.

Table 5.1: Results from training the simple CNN and VGG16 on ultrasound images and validating on ultrasound and pocket ultrasound images.

Network	Accuracy Training Ultrasound	Accuracy Validation Ultrasound	Accuracy Validation Pocket Ultrasound	Standard Deviation Validation Pocket Ultrasound
Simple CNN	0.811	0.709	0.792	0.099
VGG16	0.817	0.740	0.802	0.051

Table 5.2 shows the results from the test set on the simple CNN and the VGG16 transfer network. The table displays the accuracy on the pocket ultrasound test set and the ultrasound test set as well as the weighted accuracy, standard deviation and AUC value of the pocket ultrasound test set.
The highest accuracy, AUC value and the best ROC curve, shown in Figure 5.1, was given by the VGG16 as well as a lower standard deviation on the five models' test

Table 5.2: Results from testing the simple CNN and VGG16 on ultrasound and pocket ultrasound images.

Network	Accuracy Test Ultrasound	Accuracy Test Pocket Ultrasound	Weighted Accuracy Test Pocket Ultrasound	Standard Deviation Test Pocket Ultrasound	AUC Test Pocket Ultrasound
Simple CNN	0.674	0.817	0.794	0.095	0.91
VGG16	0.718	0.825	0.728	0.037	0.94

pocket accuracy. Although, the weighted accuracy is higher for the simple CNN thus entailing this model to have a higher accuracy on malignant images.

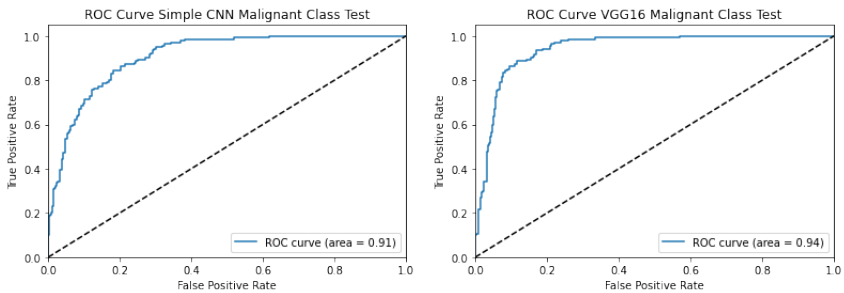

Figure 5.1: ROC curve on the pocket ultrasound test set for the simple CNN and VGG16 after training on ultrasound images.

Figure 5.2 displays the confusion matrix on the pocket ultrasound test set from part one. The figure shows the simple CNN having a higher accuracy on malignant images then the VGG16 which supports the weighted accuracy in Table 5.2 being larger for the simple CNN.

5.2 Part 2: Augmentation of ultrasound images

5.2.1 Histogram

The images used for making a histogram of the same patient and respective area are displayed in Figure 5.3 and 5.4 where Figure 5.3 shows the pocket ultrasound images and Figure 5.4 shows the ultrasound images.
The histograms corresponding to the images in Figure 5.3 and 5.4 are displayed in Figure 5.5. In this figure, the pocket ultrasound images seem to have more black pixels in both the malignant, benign and normal image. The ultrasound images on the other hand seem to have a higher frequency in the middle of the gray scale.

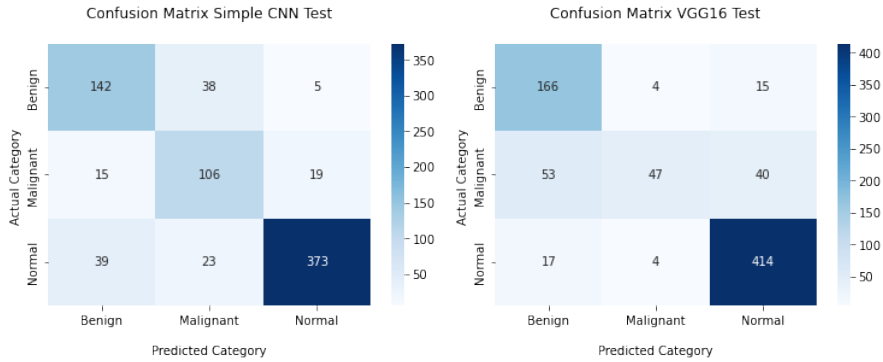

Figure 5.2: Confusion matrix on the pocket ultrasound test set for the simple CNN and VGG16 after training on ultrasound images.

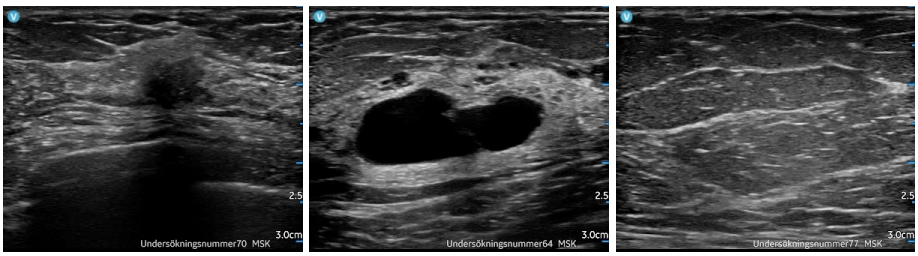

Figure 5.3: Malignant, benign and normal pocket ultrasound images.

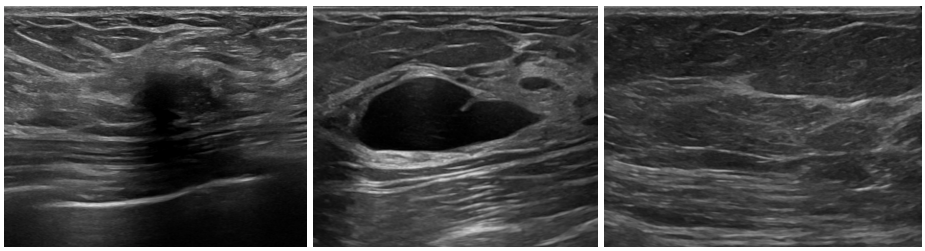

Figure 5.4: Malignant, benign and normal ultrasound images.

The histograms of pixel intensity in all ultrasound and pocket ultrasound images are showed in Figure 5.6. The figure displays the pocket ultrasound images having slightly higher intensity and twice as many black pixels then the ultrasound images. The ultrasound images are more evenly distributed and have more white pixels then the pocket ultrasound images. This conforms with the results from the histograms in Figure 5.5.

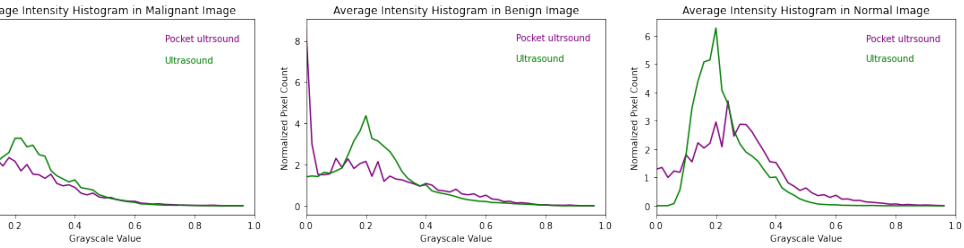

Figure 5.5: Histograms over pixel intensity in malignant, benign and normal ultrasound and pocket ultrasound image. The green line corresponds to the ultrasound image and the purple line corresponds to the pocket ultrasound image.

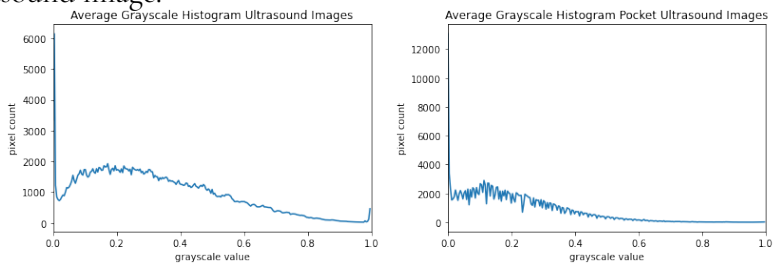

Figure 5.6: Average histogram over pixel intensity in ultrasound and pocket ultrasound images.

5.2.2 Resolution

The resolution in an ultrasound image is highly dependent on the user, in real time the doctor often changes the image view of the ultrasound probe. In order to see deeper into the mass the frequency needs to decrease for the sound waves to penetrate deeper. This will entail a decrease of resolution. Therefore it is hard to determine the difference in frequency and resolution between the pocket ultrasound and ultrasound probe. The frequency range of the pocket ultrasound on the linear probe side is 3-12MHz with a center frequency of 7.7MHz and the Logiq E9 uses the frequency range 6-15MHz [9] [38].

5.2.3 Training on Augmented Ultrasound Images

Table 5.3 shows the results from training on augmented ultrasound images which displays the simple CNN performing better on the pocket ultrasound validation set. It also has a lower standard deviation then the VGG16 model. The training accuracy and ultrasound validation accuracy is however higher for the VGG16 model.

Table 5.3: Results from training the simple CNN and VGG16 on augmented ultrasound images and validating on ultrasound and pocket ultrasound images.

Network	Accuracy Training Ultrasound	Accuracy Validation Ultrasound	Accuracy Validation Pocket Ultrasound	Standard Deviation Validation Pocket Ultrasound
Simple CNN	0.770	0.702	0.778	0.068
VGG16	0.887	0.806	0.758	0.1285

After testing, the simple CNN still performed better then the VGG16 in terms of accuracy and weighted accuracy. However, the AUC value is larger and the ROC curve in 5.7 is better for VGG16.

Table 5.4: Results from testing the simple CNN and VGG16 on augmented ultrasound and pocket ultrasound images.

Network	Accuracy Test Ultrasound	Accuracy Test Pocket Ultrasound	Weighted Accuracy Test Pocket Ultrasound	Standard Deviation Test Pocket Ultrasound	AUC Test Pocket Ultrasound
Simple CNN	0.684	0.818	0.761	0.039	0.88
VGG16	0.791	0.811	0.647	0.036	0.92

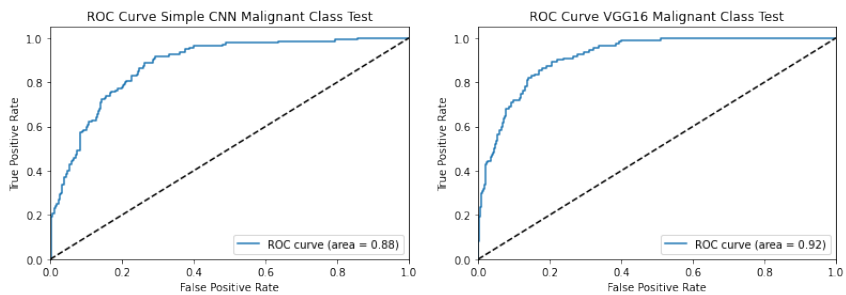

Figure 5.7: ROC curve on the pocket ultrasound test set for the simple CNN and VGG16 after training on augmented ultrasound images

The confusion matrix in Figure 5.8 displays the simple CNN performing better on the malignant class then the VGG16. However, VGG16 seem to perform better on the benign and normal class.
When comparing the results from part one and part two, the augmentation of the images does not seem to have a large impact on the generalized performance of the models. The VGG16 shows slightly better results in terms of pocket ultrasound test accuracy in part one as well as the ROC value and curve shown in Figure 5.1 and

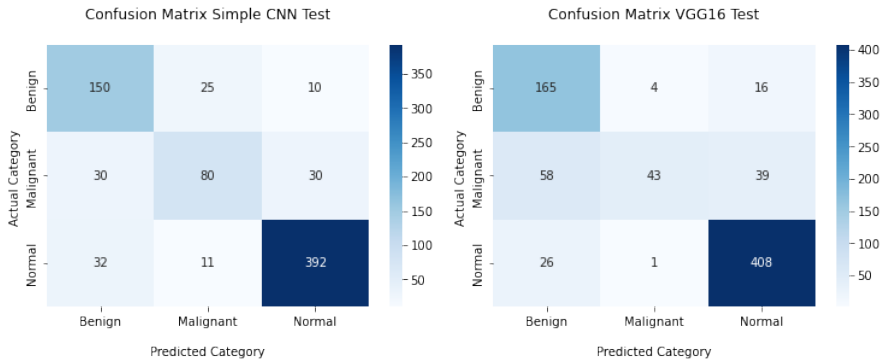

Figure 5.8: Confusion matrix on the pocket ultrasound test set for the simple CNN and VGG16 after training on augmented ultrasound images.

5.7. The augmentation does therefore not seem to help the neural networks to better classify the pocket ultrasound images.

5.3 Part 3: Training on Ultrasound and Pocket Ultrasound Images

In part three, where transfer learning was used, the generalized performance is increased noticeably. Table 5.5 displays a higher training accuracy for both VGG16 and the simple CNN than the other parts of the project. The simple CNN is also slightly higher in training accuracy and pocket validation accuracy then the VGG16 models.

Table 5.5: Results from training and validating the simple CNN and VGG16 on both ultrasound and pocket ultrasound images.

Network	Accuracy Training Ultrasound	Accuracy Validation Ultrasound	Accuracy Validation Pocket Ultrasound	Standard Deviation Validation Pocket Ultrasound
Simple CNN	0.927	0.511	0.797	0.082
VGG16	**0.890**	**0.673**	**0.787**	**0.079**

After testing, the VGG16 performs noticeably better then the simple CNN on all accounts, as Figure 5.6 displays. The ROC curve and AUC value are also better for the VGG16 then for the simple CNN as can be viewed in Figure 5.9.
The VGG16 also performs good on all three classes as can be seen in the right confusion matrix in Figure 5.10 whereas the simple CNN has trouble classifying malignant and normal images as benign, as can be seen in the left confusion matrix.

Table 5.6: Results from testing the simple CNN and VGG16 on ultrasound and pocket ultrasound images.

Network	Accuracy Test Ultrasound	Accuracy Test Pocket Ultrasound	Weighted Accuracy Test Pocket Ultrasound	Standard Deviation Test Pocket Ultrasound	AUC Test Pocket Ultrasound
Simple CNN	0.478	0.795	0.777	0.070	0.89
VGG16	**0.662**	**0.868**	**0.849**	**0.04**	**0.93**

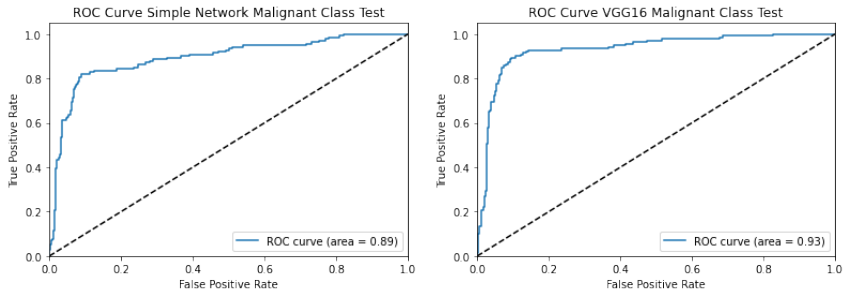

Figure 5.9: ROC curve and AUC value on the pocket ultrasound test set for the simple CNN and VGG16 after training on ultrasound and pocket ultrasound images.

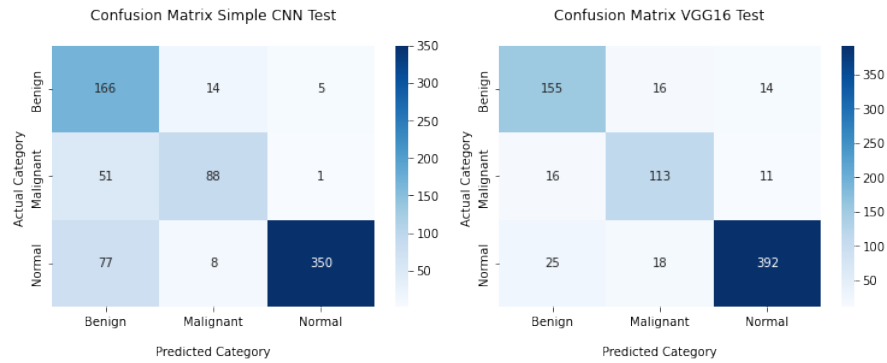

Figure 5.10: Confusion matrix on the pocket ultrasound test set for the simple CNN and VGG16 after training on ultrasound images and pocket ultrasound images.

When comparing the results from part three with the first two parts of the project, the VGG16 in part three seems to have the best overall performance. The pocket ultrasound test accuracy and weighted accuracy in Table 5.6 is better then in the other parts, as viewed in Table 5.4 and 5.2. The ROC curve and value from Figure 5.9 is better

or approximately the same compared to the one in part one and two shown in Figure 5.1 and 5.7.

5.4 Part 4: Training on Pocket Ultrasound Images

The last part of the project where the networks were trained on only pocket ultrasound images, the validation accuracy is quite low for both pocket ultrasound and ultrasound images.

Table 5.7: Results from training the simple CNN and VGG16 on pocket ultrasound images and validating on ultrasound and pocket ultrasound images.

Network	Accuracy Training Ultrasound	Accuracy Validation Ultrasound	Accuracy Validation Pocket Ultrasound	Standard Deviation Validation Pocket Ultrasound
Simple CNN	0.808	0.315	0.741	0.041
VGG16	0.935	0.528	0.768	0.049

After testing, the pocket validation accuracy and weighted accuracy is better for the VGG16 then the simple CNN, as Table 5.8 displays. The standard deviation of the five VGG16 models is also noticeably lower then the five simple CNN models.

Table 5.8: Results from testing the simple CNN and VGG16 on ultrasound and pocket ultrasound images.

Network	Accuracy Test Ultrasound	Accuracy Test Pocket Ultrasound	Weighted Accuracy Test Pocket Ultrasound	Standard Deviation Test Pocket Ultrasound	AUC Test Pocket Ultrasound
Simple CNN	0.310	0.797	0.682	0.114	0.85
VGG16	0.533	0.825	0.814	0.027	0.90

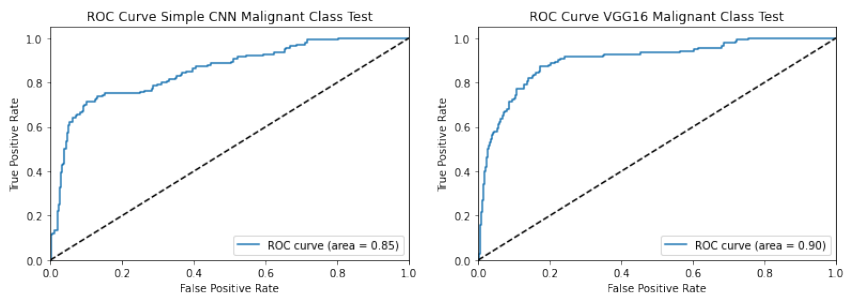

Figure 5.11: ROC curve and AUC value on the pocket ultrasound test set for the simple CNN and VGG16 after training on pocket ultrasound images.

The confusion matrix in Figure 5.12 display the VGG16 being good at classifying all classes whereas the simple CNN have trouble classifying malignant images and benign images as normal.

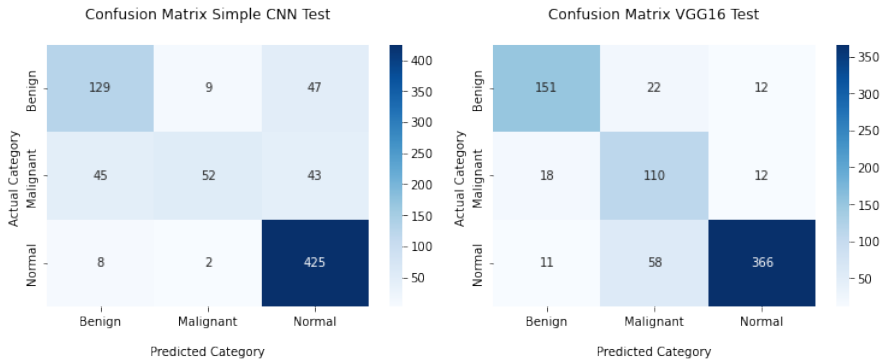

Figure 5.12: Confusion matrix on the pocket ultrasound test set for the simple CNN and VGG16 after training on pocket ultrasound images.

Comparing the results from part four to the other parts of the thesis, the pocket validation accuracy, in Table 5.8, is approximately the same as in part one and two. The standard deviation of the five VGG16 models, shown in Table 5.8, is lower than in any other part of the project. Furthermore, the ROC curve and AUC value displayed in 5.11 are worse then in the other parts of the project.

Chapter 6
Discussion

6.1 Discussion of Performance

From the results in chapter 5 one can see that training on ultrasound images helps to classify the pocket ultrasound images but it seems the differences of the two types of ultrasound images is too great to get a good accuracy by solely training on ultrasound images. Training on only pocket ultrasound images is not optimal with the amount of data available but considering the accuracy given, it could be the best approach if more data is collected. The best approach from this thesis, given the amount of data, would reasonably be to train the model partly on ultrasound images and then optimizing the model features on pocket ultrasound data, i.e. using transfer learning. When looking at the results, the VGG16 model in part three seems to perform best on the pocket ultrasound data which agrees with the theory mentioned above.

6.2 Evaluation of Data sets

6.2.1 Egyptian Data Set

Many images in the Egyptian data set contains annotations of findings and the size of findings. This can affect the result of the classification since only benign and malignant images can have annotations, there is nothing to annotate in a normal image. The algorithm can therefore learn that annotations and marks in the image is a feature for malignant or benign images. The images tested by the algorithm might not have

annotations and could therefore be wrongly classified as normal. However since the ultrasound data set consists of three different data sets where two of them are lacking annotations this will probably not affect the overall result.

6.2.2 Dutch Data Set

The online data set only contains images of benign and malignant findings. The normal images was created by cropping normal parts of the existing images. The data set is therefore not balanced, only 1/6 of the images are normal and these images are more zoomed in than the rest. This could lead to the algorithms finding features based on the zoom of the image and not the actual important features. However since the other data sets contain normal images the algorithm should not choose features of the normal class based solely on the cropped normal images in the Dutch data set.

6.2.3 Swedish Data Set

Some of the patients the pocket ultrasound images were taken from also had ultrasound images that were used in the project as well. They were mainly used to make the pixel intensity histograms but they were also included in the joined ultrasound data set during training. There is therefore a possibility that a pocket ultrasound image in the validation set is similar to an image in the ultrasound training data set since the images could be taken on the same patient and finding. This can entail a better result on the validation due to the algorithm recognizing an image from the training data set. It could also entail the algorithm to not learn and find important features from the similar image since it already knows the answer. The pocket ultrasound patients' images in the test set is however not present in the ultrasound training set. Therefore the effect of the matter is unessential to the results produced since the effect only occurs to the validation data. It can however entail the validation data to be a worse representation of the test data and therefore the best model selected from the validation data may not be the best model for the test data.

6.2.4 Pocket Ultrasound Data Set

Since multiple images were taken on the same patient, the data sets were divided so that one patients' images would not exist in both training and testing sets. This avoids bias problems where the algorithm has already seen similar data in the training set. The method was also used when creating the validation sets since they should represent the test data. Since the data set is so small and many images were taken from the same patient, some of the pocket validation sets only consists of one patient in the

malignant class. The validation results will therefore be highly effected by how difficult that specific finding is. Some images were also taken from patients with unusual findings, such as scar tissue from mastectomy or inflammations and since the data set is small these images could be hard for the algorithm to classify due to lack of such images in the training set. When training the model, this was the case on one of the five validation sets which constantly on all parts of the project gave significantly lower results than the other validation sets due to it randomly happen to have hard patients to classify. The overall validation results would therefore be lower since it is calculated by taking an average over all validation results. This could explain why the validation results are lower then the test result in almost all parts of the project. The problem would probably be solved with the use of more pocket ultrasound data. The collection of large data sets of pocket ultrasound images is cumbersome, timewise, since it is not part of routine health care and requires informed consent. Furthermore, the pocket ultrasound images were classified by one radiologist. To be more sure of the true class of the images, two radiologist should preferably classify the images. However, if uncertainty existed regarding the diagnosis of a finding, samples are taken. In this thesis, the classes of these 'uncertain' images were determined after the sample results were known.

6.3 Limitations

The created algorithm is made for classifying pocket ultrasound images and is optimized to do so meaning it is not trained to classify ordinary ultrasound images or pocket ultrasound images in other parts of the body then the breasts. The pocket ultrasound images are taken from the Vscan Air probe manufactured by GE, meaning the results are optimized from this device and is not surely to perform the same on images from similar devices from different manufacturers.

6.4 Future Development

6.4.1 Double K-fold Cross Validation Loops

In this project, a situation were the validation performance was lower then the test performance appeared. The validation data is divided, as mentioned before, into five different parts. The test data however is the same in all five models thus entailing the images in the test set to be highly influential. Since the data set is small, there is a possibility that easily classified images randomly happen to end up in the test set. This

will give a higher test performance which may not agree on new data. A way to avoid this would be to implement a double k-fold cross validation loop where k different test sets are produced in the same way as the k validation sets. This method was however not optimal in this project due to lack of data but could be a future development of this thesis if more data is collected. Figure 6.1 describes the double k-fold cross validation method.

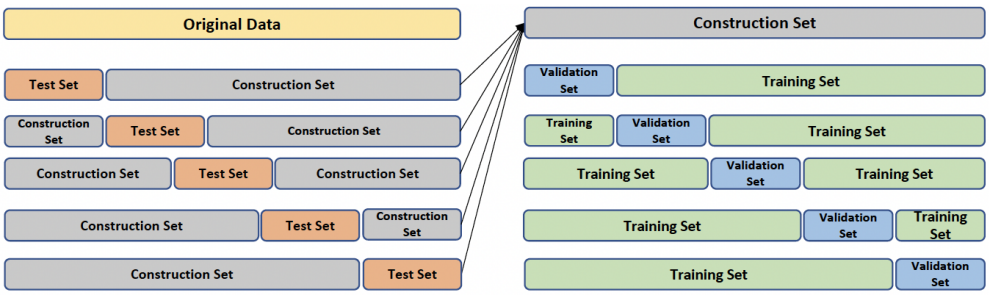

Figure 6.1: Illustration of the double k-fold cross validation method. The image displays an example with 5 folds.

6.4.2 Voting

In this thesis the test set was predicted by all five models and all their predictions were used to calculate the evaluation metrics in the result chapter. If one model is bad at classifying a certain image its' prediction will influence the result. One way to avoid this is to implement *voting* of the models. All five models would then classify one image and the most common classification from the five models would be the predicted class of that image.

6.4.3 Heatmaps /Grad-CAM

The results from this thesis does not portray what in the images that makes the algorithm classify them a certain way. To understand what features the algorithm finds fitting to a certain class, Grad-CAM can be used. The grad-CAM creates a *heatmap* over a predicted image and displays in a colour scheme what part of the image that is an important feature in the classification [50]. This would entail better understanding of the model and possibly also entail better solutions and classification results.

6.4.4 Evaluation of the needed amount of data

It would be useful to get an understanding of how much pocket ultrasound data is needed to produce an acceptable accuracy. One way to achieve this is by training, validating and testing the models on only ultrasound data to get an idea of how good the model could be with more data available. The number of ultrasound images would then be reduced to the same amount of images as the pocket ultrasound data set has. These images would be used in training to get an idea of if the complexity of the problem is the same for pocket ultrasound and ultrasound images. The amount of ultrasound images would then be increased step wise and training would occur in every step. The accuracy can then be plotted against the number of images to get an idea of how many pocket ultrasound images would be needed to attain a certain accuracy.

Chapter 7
Conclusion

The results obtained in this thesis suggest that using ultrasound images during training helps to successfully classify malignant, benign and normal tissue in pocket ultrasound images. However, the two types of ultrasound images differ to some extent, thus limiting higher performance. Furthermore, using augmented ultrasound images during training does not affect the overall performance of the models.

Due to limited data, training solely on pocket ultrasound images is not optimal but could be a good method to use with more data available. The best model obtained in this thesis was created by using ultrasound and pocket ultrasound images in transfer learning. The test accuracy from this model was 86.8% and the AUC value was 0.93.

From the results in this thesis it seems possible to use deep learning in order to classify breast cancer in pocket ultrasound images. This method could therefore have a great impact, especially in low income countries. After collecting more data and the model is perfected to perform at a high accuracy, the algorithm together with a pocket ultrasound probe could possibly provide breast diagnostics in low-resource settings. It could thereby help reduce the mortality of breast cancer and avoid unnecessary mastectomies of benign breast findings.

References

[1] Khaled Fahmy Al-Dhabyani, Gomaa. Dataset of breast ultrasound images. *Data in Brief*, 2020.

[2] Arvidsson. Applications of deep learning in medical image analysis, grading of prostate cancer and detection of coronary artery disease. 2021.

[3] Alemayehu et al. Bawoke, Kejela. Experience with modified radical mastectomy in a low-income country: a multi-center prospective observational study. *BMC Surgery*, 2021.

[4] Cormack et al. Berg, Blume. Combined screening with ultrasound and mammography vs mammography alone in women at elevated risk of breast cancer. *JAMA*, 2008.

[5] Bröstcancerförbundet. Täta bröst. `https://brostcancerforbundet.se/om-brostcancer/diagnostik/tata-brost/` [Data accessed: 2022-04-22].

[6] Chollet. *Deep learning with python*. Manning, 20 Baldwin Road, Shelter Island, 2017.

[7] GE General Electric Company. Vscan air. `https://vscan.rocks/product/vscanair?utm_source=Google&utm_medium=Paid%20Search&utm_campaign=EMEA-22-01-US-Digital-HH-Vscan-Rocks-Site-Paid-Search&gclid=Cj0KCQjwma6TBhDIARIsAOKuANxr19LdFdize2qZRC3lphXqiO_a53OoZQw4c_oc27swyJ4As7jiafMaAnOpEALw_wcB` [Data accessed: 2022-04-30].

[8] GE General Electric Company. Vscan air. https://vscan.rocks/product/vscanair [Data accessed: 2022-04-23].

[9] GE General Electric Company. Vscan air data sheet. https://www.gehealthcare.com/-/media/A66296638B1D4D4EA4B9411857A38066.pdf [Data accessed: 2022-05-2].

[10] Ekwueme Coughlin. Breast cancer as a global health concern. *Cancer Epidemiology*, 2009.

[11] Dinesh. Cnn vs mlp for image classification. *Analytics Vidhya*, 2019.

[12] Freudenrich. How ultrasound works. https://www.physics.utoronto.ca/~jharlow/teaching/phy138_0708/lec04/ultrasoundx.htm [Data accessed: 2022-04-22].

[13] Chen et al. Georgiou, Liu. A survey of traditional and deep learning-based feature descriptors for high dimensional data in computer vision. *International Journal of Multimedia Information Retrieval*, 2020.

[14] Ginsburg. Breast and cervical cancer control in low and middle-income countries: Human rights meet sound health policy. *Journal of Cancer Policy*, 2013.

[15] Courville Goodfellow, Bengio. *Deep Learning*. The MIT press, Cambridge, Massachusetts, 2016.

[16] Jin. Cancer screening: Benefits and harms. *JAMA*, 2014.

[17] Safaraliev Cherukuri Zgheib Kamalov, Nazir. Comparative analysis of activation functions in neural networks. 2021.

[18] Ramkull Karlsson. Machine learning algorithm for classification of breast ultrasound images. 2021.

[19] IBM Kavlakoglu, Program Manager. Ai vs. machine learning vs. deep learning vs. neural networks: What's the difference? https://www.ibm.com/cloud/blog/ai-vs-machine-learning-vs-deep-learning-vs-neural-networks [Data accessed: 2022-05-13].

[20] Comulada et al. Kelly, Dean. Breast cancer detection using automated whole breast ultrasound and mammography in radiographically dense breasts. *European Radiology*, 2009.

[21] Keras. About keras. https://keras.io/about/ [Data accessed: 2022-05-10].

[22] Keras. Inceptionresnetv2. https://keras.io/api/applications/inceptionresnetv2/ [Data accessed: 2022-05-10].

[23] Keras. Inceptionv3. https://keras.io/api/applications/inceptionv3/ [Data accessed: 2022-05-10].

[24] Keras. Vgg16 and vgg19. https://keras.io/api/applications/vgg/ [Data accessed: 2022-05-10].

[25] Keras. Xception. https://keras.io/api/applications/xception/ [Data accessed: 2022-05-10].

[26] LBNMedical. How much does a mammogram machine cost – easy overview. https://lbnmedical.com/how-much-does-a-mammogram-machine-cost/ [Data accessed: 2022-04-30].

[27] Scikit learn developers (BSD License). sklearn.utils.class_weight.compute_class_weight. https://scikit-learn.org/stable/modules/generated/sklearn.utils.class_weight.compute_class_weight.html [Data accessed: 2022-04-30].

[28] Creative Commons License. Attribution 4.0 international (cc by 4.0). https://creativecommons.org/licenses/by/4.0/ [Data accessed: 2022-05-19].

[29] Malik. Neural networks bias and weights. *FinTechExplained*, 2019.

[30] Brock Shulman Martei, Pace. Breast cancer in low- and middle-income countries. *Clinics in Laboratory Medicine*, 2018.

[31] El-Feky Morovati. Acoustic enhancement. 2021.

[32] NumPy. Numpy documentation. https://numpy.org/doc/ [Data accessed: 2022-05-10].

[33] Ogoti. Why python is good for machine learning. https://www.section.io/engineering-education/why-python-is-good-for-machine-learning/ [Data accessed: 2022-05-2].

[34] Edén Ohlsson. *Lecture notes on Introduction to Artificial Neural Networks and Deep Learning*. Media-tryck, Department of Astronomy and Theoretical Physics, Lund University, 2020.

[35] Bugeza et al. Okello, Kisembo. Breast cancer detection using sonography in women with mammographically dense breasts. 2014.

[36] World Health Organization. Breast. `https://gco.iarc.fr/today/data/factsheets/cancers/20-Breast-fact-sheet.pdf` [Data accessed: 2022-02-20].

[37] Pillow. Imageenhance module. `https://pillow.readthedocs.io/en/stable/reference/ImageEnhance.html` [Data accessed: 2022-05-10].

[38] Kristina Lång; Radiologist at Unilabs Mammography Unit at Skåne University Hospital.

[39] Scikit-image. random_noise. `https://scikit-image.org/docs/dev/api/skimage.util.html#skimage.util.random_noise` [Data accessed: 2022-05-10].

[40] scikit learn. scikit-learn, machine learning in python. `https://scikit-learn.org/stable/` [Data accessed: 2022-05-10].

[41] Cancer Research UK Scowcroft. Why are breast cancer rates increasing? `https://news.cancerresearchuk.org/2011/02/04/why-are-breast-cancer-rates-increasing/` [Data accessed: 2022-05-19].

[42] Lundervold Selvikvåg Lundervold. An overview of deep learning in medical imaging focusing on mri. *Zeitschrift für Medizinische Physik*, 2019.

[43] Sharma. Understanding transfer learning for deep learning. `https://www.analyticsvidhya.com/blog/2021/10/understanding-transfer-learning-for-deep-learning/` [Data accessed: 2022-04-26].

[44] Yonso Skalski. Acoustic shadowing. 2020.

[45] Stanford University Stanford Vision Lab. Imagenet. `https://image-net.org/index` [Data accessed: 2022-05-13].

[46] Ng Swanevelder. Resolution in ultrasound imaging. *Continuing Education in Anaesthesia Critical Care Pain*, 2011.

[47] Tensorflow.org. Classification on imbalanced data. `https://www.tensorflow.org/tutorials/structured_data/imbalanced_data#class_weights` [Data accessed: 2022-04-30].

[48] Mufunda Tesfamariam, Gebremichael. Breast cancer clinicopathological presentation, gravity and challenges in eritrea, east africa: Management practice in a resource-poor setting. *South African Medical Journal*, 2013.

[49] Bouvry Pinel Leprévost Thanapol, Lavangnananda. Reducing overfitting and improving generalization in training convolutional neural network (cnn) under limited sample sizes in image recognition. 2020.

[50] Park Hwang et al. Yoon, Han. Optical coherence tomography-based deep-learning model for detecting central serous chorioretinopathy. *Scientific Reports*, 2020.

[51] Özkurt Şen. Convolutional neural network hyperparameter tuning with adam optimizer for ecg classification. 2020.