LUND UNIVERSITY

Deep Learning for Breast Cancer Detection in Ultrasound Imaging

Classification, Managing Data Scarcity and Detecting Out-of-Distribution Samples

Karlsson, Jennie

2024

– CENTRUM SCIENTIARUM MATHEMATICARUM –

Deep Learning for Breast Cancer Detection in Ultrasound Imaging

Classification, Managing Data Scarcity and Detecting Out-of-Distribution Samples

JENNIE KARLSSON

Lund University
Faculty of Engineering
Centre for Mathematical Sciences
Mathematics

Deep Learning for Breast Cancer Detection in Ultrasound Imaging

Classification, Managing Data Scarcity and Detecting Out-of-Distribution Samples

by Jennie Karlsson

LUND
UNIVERSITY

LICENTIATE THESIS

which, by due permission of the Faculty of Engineering at Lund University, will be publicly defended on Friday 6th of September, 2024, at 14:00 in lecture hall MH:309A.

Thesis advisors:
A. Heyden, N.C. Overgaard, K. Åström, I. Arvidsson, K. Lång

Faculty opponent:
P. Edén

Organization
LUND UNIVERSITY
Centre for Mathematical Sciences
Box 118
SE–221 00 LUND
Sweden

Document name
Licentiate thesis

Date of presentation
2024-09-06

Sponsoring organization
eSSENCE, AIDA Vinnova Grant 2017-02447 and 2021-01420

Author(s)
Jennie Karlsson

Title and subtitle
Deep Learning for Breast Cancer Detection in Ultrasound Imaging – Classification, Managing Data Scarcity and Detecting Out-of-Distribution Samples

Abstract
Breast cancer has a profound affect on society. The survival for women in low- and middle-income countries (LMICs) is poor compared to in high-income countries (HICs). The lack of timely diagnosis is one of the main factors contributing to the poor outcomes for women in LMICs. Point-of-care ultrasound (POCUS) combined with a deep learning (DL) classification network could potentially be a suitable support tool for breast cancer detection in LMICs.

There are different ways of designing a classification network. Convolutional neural networks (CNNs) are widely used for image classification tasks and are becoming regular in medical applications. To train a DL classification network data is needed. Medical data can be difficult to acquire due to many different reasons, one of them being ethical approvals. The availability of breast POCUS data is limited. However, the access to standard ultrasound (US) images is greater. There are different methods to expand a data set, a very common one is the use of data augmentation. Another interesting method is the cycle-consistent adversarial network (CycleGAN). This network is trained to transform an image of one domain into another domain. Thus, standard US images could be transformed into the domain of POCUS imaging, generating more POCUS samples without the need of collecting the POCUS images in the clinic. Further, it is crucial to assure the classification network to be trustworthy. Images which are out-of-distribution (OOD) should be detected and no prediction by the network should be made. OOD samples in breast US imaging includes; images of poor quality, images capturing other structures than breast tissue and images showing rare lesions. For the third case, rare lesions, the sample is in-distribution (ID). However, since the lesion is rare and might not be covered within the knowledge of the classification network it should still not be predicted. In such cases, the network's uncertainty of the prediction can be used in order to decide whether a safe prediction can be made or not. This is called uncertainty quantification.

This thesis includes four papers covering different steps towards implementing a breast POCUS classification network. The first paper is focusing on classification of standard US images. This is a first step, showing the potential of using DL for breast cancer classification. In the second paper the classification network is modified to classify POCUS data. In order to expand the POCUS data, augmentation and CycleGAN are used. The third paper cover OOD detection and methods such as energy score and deep ensembles are studied. Finally, uncertainty quantification is covered in the fourth paper. Excluding samples with high uncertainty did improve the classification result. The papers included in this thesis show that there is potential of using DL in a trustworthy way for breast cancer detection in LMICs.

Key words
Breast cancer detection; Low-resource settings; Point-of-care ultrasound imaging; Deep learning; Image domain shift; Out-of-distribution detection; Uncertainty quantification

Classification system and/or index terms (if any)

Supplementary bibliographical information

Language
English

ISSN and key title
1404-028X

ISBN
978-91-8104-157-6 (print)
978-91-8104-158-3 (electronic)

Recipient's notes

Number of pages
xii+100

Price

Security classification

Deep Learning for Breast Cancer Detection in Ultrasound Imaging

Deep Learning for Breast Cancer Detection in Ultrasound Imaging

Classification, Managing Data Scarcity and Detecting Out-of-Distribution Samples

by Jennie Karlsson

LUND
UNIVERSITY

Centre for Mathematical Sciences
Lund University
Box 118
SE-221 00 Lund
Sweden

www.maths.lu.se

Licentiate Thesis in Mathematical Sciences 2024:2
ISSN: 1404-028X
ISBN: 978-91-8104-157-6 (print)
ISBN: 978-91-8104-158-3 (electronic)
LUTFMA-2046-2024

Printed in Sweden by Media-Tryck, Lund University, Lund 2024

Printed matter
3041 0903

Media-Tryck is an Nordic Swan Ecolabel
certified provider of printed material.
Read more about our environmental
work at www.mediatryck.lu.se

MADE IN SWEDEN

Abstract

Breast cancer has a profound affect on society. The survival for women in low- and middle-income countries (LMICs) is poor compared to in high-income countries (HICs). The lack of timely diagnosis is one of the main factors contributing to the poor outcomes for women in LMICs. Point-of-care ultrasound (POCUS) combined with a deep learning (DL) classification network could potentially be a suitable support tool for breast cancer detection in LMICs.

There are different ways of designing a classification network. Convolutional neural networks (CNNs) are widely used for image classification tasks and are becoming regular in medical applications. To train a DL classification network data is needed. Medical data can be difficult to acquire due to many different reasons, one of them being ethical approvals. The availability of breast POCUS data is limited. However, the access to standard ultrasound (US) images is greater. There are different methods to expand a data set, a very common one is the use of data augmentation. Another interesting method is the cycle-consistent adversarial network (CycleGAN). This network is trained to transform an image of one domain into another domain. Thus, standard US images could be transformed into the domain of POCUS imaging, generating more POCUS samples without the need of collecting the POCUS images in the clinic. Further, it is crucial to assure the classification network to be trustworthy. Images which are out-of-distribution (OOD) should be detected and no prediction by the network should be made. OOD samples in breast US imaging includes; images of poor quality, images capturing other structures than breast tissue and images showing rare lesions. For the third case, rare lesions, the sample is in-distribution (ID). However, since the lesion is rare and might not be covered within the knowledge of the classification network it should still not be predicted. In such cases, the network's uncertainty of the prediction can be used in order to decide whether a safe prediction can be made or not. This is called uncertainty quantification.

This thesis includes four papers covering different steps towards implementing a breast POCUS classification network. The first paper is focusing on classification of standard US images. This is a first step, showing the potential of using DL for breast cancer classification. In the second paper the classification network is modified to classify POCUS data. In order to expand the POCUS data, augmentation and CycleGAN are used. The third paper cover OOD detection and methods such as energy score and deep ensembles are studied. Finally, uncertainty quantification is covered in the fourth paper. Excluding samples with high uncertainty did improve the classification result. The papers included in this thesis show that there is potential of using DL in a trustworthy way for breast cancer detection in LMICs.

List of Publications

This thesis is based on the following publications,

Main papers

I **Machine Learning Algorithm for Classification of Breast Ultrasound Images**

J. Karlsson, J. Ramkull, I. Arvidsson, A. Heyden, K. Åström, N.C. Overgaard and K. Lång
Proceedings of the International Society for Optics and Photonics (SPIE), Medical Imaging: Computer-Aided Diagnosis, 2022 [13].

II **Classification of Point-of-Care Ultrasound in Breast Imaging Using Deep Learning**

J. Karlsson, I. Arvidsson, F. Sahlin, K. Åström, N.C. Overgaard, K. Lång and A. Heyden
Proceedings of the International Society for Optics and Photonics (SPIE), Medical Imaging: Computer-Aided Diagnosis, 2023 [14].

III **Towards Out-of-Distribution Detection for Breast Cancer Classification in Point-of-Care Ultrasound Imaging**

J. Karlsson, M. Wodrich, N.C. Overgaard, F. Sahlin, K. Lång, A. Heyden, I. Arvidsson
Proceedings of the 27th International Conference on Pattern Recognition (ICPR), 2024 (accepted and to appear in) [15].

IV **Trustworthiness for Deep Learning Based Breast Cancer Detection Using Point-of-Care Ultrasound Imaging in Low-Resource Settings**

M. Wodrich, **J. Karlsson**, K. Lång, I. Arvidsson
Proceedings of the MICCAI meets Africa Workshop at the 27th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) , 2024 (accepted and to appear in) [34].

Author's Contributions

I This paper is based on the master's thesis of JK and JR. KL had the original idea for the project. The implementation of code was done by me and JR. Most of the paper was written by me with input from the co-authors.

II The original idea came from me. I did the implementation and wrote the paper, with input from all the co-authors.

III The original idea came from me and IA. The code was implemented by me and MW. The paper was mostly written by me, with revision by the co-authors.

IV The paper evolved form the master's thesis written by MW, supervised by me and IA. MW did the implementations and wrote most of the paper with revision by the co-authors, including me.

Acknowledgements

I want to start with expressing gratitude to my main supervisor Anders Heyden and my co-supervisors Niels-Christian Overgaard and Kalle Åström. Thank you for your support, encouragement and for valuable discussions and input. I would also like to express gratitude to my co-supervisor Ida Arvidsson. Thank you for supporting me, answering all my questions and being an inspiring role model. Further, I want to thank my co-supervisor Kristina Lång for giving me the opportunity to be a part of this interesting project and for guiding me in the complex field of medicine. To all my amazing colleagues at the centre for mathematical sciences, thank you for making the workplace fun and supporting. I also want to thank all of the co-authors and project members for good collaborations and valuable discussions.

Finally I want to express gratitude to my family. Especially, Mamma, for your unconditional support, and Jari, for always encouraging me to follow my goals and making mathematics fun. I also want to thank Felix, for giving me the confidence needed to start this PhD position.

Funding

This work has been supported by strategic research area eSSENCE. It has also obtained grant support by Analytic Imaging Diagnostics Arena (AIDA), Vinnova Grant 2017-02447 and 2021-01420.

Contents

Chapter 1

Introduction

Approximately one out of six deaths worldwide is caused by cancer [3], giving it a profound impact on society. This impact is further corroborated by recent studies estimating nearly one million children became maternal orphans during 2020 due to losing their mothers to cancer, of which breast cancer was the largest contributing cause of maternal death [9].

Cancer diseases are characterized by uncontrolled cell growth. This uncontrolled growth could spread to surrounding tissues in the body, potentially having a fatal impact. From statistics concluded in 2022, lung cancer was the cancer disease with both the highest incidence and mortality rate for women and men combined. However, for women breast cancer has both the highest incidence and mortality rate compared to other types of cancers. The number of new cases in 2022 for the five most common cancers can be observed in Table 1.1 [3].

Breast cancer mortality for women in high-income countries (HICs) has decreased remarkably during the past three decades. One of the key factors is the implementation of screening programs with mammography [29]. Meanwhile, in low- and middle-income countries (LMICs) breast cancer survival is poor compared to in HICs, showing a significant health inequity. The five year survival of breast cancer in HICs is exceeding 90 percent. In contrast, the survival rates in India and South Africa are at 66 and 40 percent respectively [33].

Table 1.1: Number of new cases in 2022 for the five most common cancer diseases.

Cancer Disease	Number of New Cases
Lung	2,480,301
Female breast	2,308,897
Colorectum	1,926,118
Prostate	1,466,680
Stomach	968,350

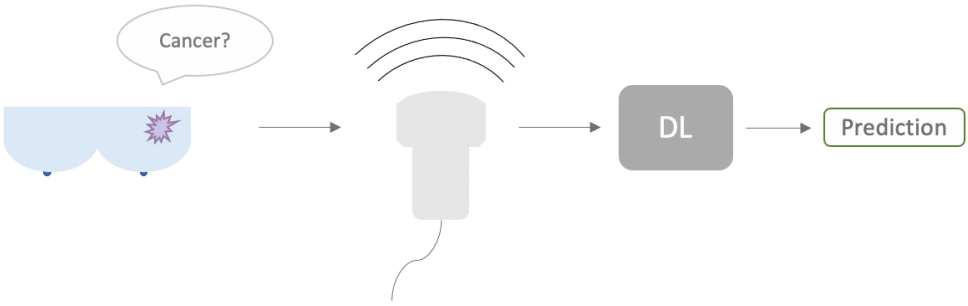

Figure 1.1: The first step in the pipeline of the suggested support tool for breast cancer detection is a women seeking healthcare due to an unknown lump in the breast. This is followed by an examination with POCUS which is analyzed by DL in order to give a prediction.

One of the main reasons for the low survival rate in LMICs is the poor access to timely diagnosis due to lack of diagnostic tools and expertise.

Mastectomy is one of the most common ways of treating breast cancer in LMICs [32]. This type of surgery is related to a higher morbidity and is mutilating compared to breast-conserving surgery. Late-stage diagnosis is a contributing factor to mastectomy being more common. Finding a suitable diagnostic tool would enable earlier diagnosis, thus avoiding unnecessary mastectomies.

Mammography is widely used in screening programs in HICs and have been shown to reduce breast cancer mortality. However, it comes with the drawbacks of being expensive and requiring a well organized health-care infrastructure, making it less suitable for LMICs. Another common method for breast examination is standard ultrasound (US), which is the type of high-end US used in the hospitals today. In recent years the quality of point-of-care ultrasound (POCUS), a pocket-sized US device, has increased. An advantage with this type of US is the low cost compared to both mammography and standard (high-end) US imaging. Several of the POCUS systems at the market today consists of a single US probe connected to a smartphone through Bluetooth or cable. This cost efficient and compact examination tool has potential of being used in LMICs for breast cancer diagnosis.

Breast US images need to be interpreted by an expert, typically a radiologist. However, in many LMICs there is a lack of radiologists. During the past years, deep learning (DL) has gained large interest in medical applications, and it has proven to generate impressive results [28]. Using a POCUS device combined with a DL classifier could potentially be an implementable support tool for breast cancer detection even in LMICs. Further, this could also be suitable in rural parts of HICs. An illustration showing the suggested pipeline for the support tool is shown in Figure 1.1.

The first paper covered in this work is a first step towards implementing a breast cancer

classifier for standard US images. Further, training DL requires data and collecting medical data from patients often requires ethical approvals and consent from the patients, making it difficult to collect. In the second paper the classifier is adapted for POCUS, and techniques for increasing the amount of training data are implemented. Moreover, it is important for the classifier to discard data samples which are out-of-distribution (OOD), i.e. outside of the knowledge of the classifier. This is covered in the third paper, where OOD detection is used prior to classification. Finally, the classifier should be trustworthy. If the classifier has high uncertainty the prediction should not be trusted. Improving the trustworthiness with uncertainty quantification (UQ) is studied in the fourth paper.

Prior to the presentation of the four papers, following chapters will be presented: Chapter 2 includes some essential machine learning (ML) techniques and Chapter 3 covers breast US data. This is followed by Chapter 4 covering breast cancer classification. In Chapter 5 we will address scarcity of medical data and how to handle it and in Chapter 6 we show how OOD detection and UQ are important in order to implement a trustworthy classifier. Finally Chapter 7 will conclude the most important findings from the papers.

Chapter 2

Machine Learning Techniques

This chapter will cover some key concepts in ML, starting with methods such as convolutional neural network (CNN) and generative adversarial network (GAN). Further, the usage of data in ML will be described. This will focus on how to split data into different sets, and how this can be done iteratively with cross-validation. Finally, some common metrics used for evaluation will be mentioned.

2.1 Convolutional Neural Networks

DL is a subset of ML that includes deep neural networks (DNNs). These networks are characterized by multiple layers with adjustable weights, which are optimized during training. One of the most popular types of DNNs is the CNN, an artificial neural network (ANN) containing at least one convolutional layer. This type of network is widely used for tasks within image analysis.

The CNN takes structured data such as images as input. In the image case, the first two dimensions contains spatial information and the final dimension contains a feature vector for each pixel. If e.g. the network inputs an RGB image, the spatial dimension will be represented by the image size. The values from each color channel would make up the vectors for the final dimension. For a grayscale image the final dimension will contain one value for each pixel.

The convolutional layer consists of a kernel with weights which slides over the input creating a feature map, allowing the network to use the spatial information. Figure 2.1 illustrates the mechanism of the convolutional operation for an input of size 5×5 with a kernel of size 3×3. First the kernel is flipped in both vertical and horizontal directions. Then the

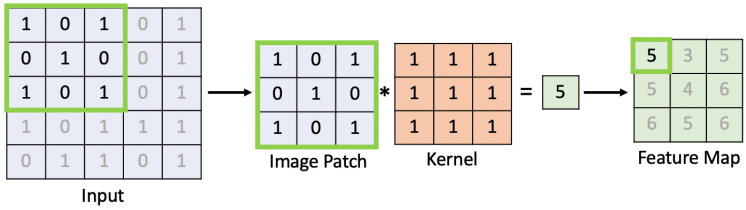

Figure 2.1: A convolution is perform on an input of size 5×5. The kernel is of size 3×3 and the stride is set to one. Each image patch of the input is cross-correlated with the flipped kernel. This results in a new feature map of size 3×3.

cross-correlation is computed between the kernel and each image patch. In this case the stride is set to one, meaning that the kernel will slide one step between each operation, covering in total nine different image patches.

It is common to use pooling layers in order to reduce the resolution, making the feature maps smaller. There are different types of pooling methods. Commonly pooling by max or average are used.

An example of a CNN architecture can be seen in Figure 2.2. The CNN inputs an RGB image into multiple convolutional layers followed by fully connected layers. The output of the network consists of three nodes, which could be predictions for three classes.

Layers in a neural network which are neither input or output are called hidden layers. For the nodes in the hidden and output layers, activation functions are used in order to find non-linear relations. A common activation is the sigmoid function, defined as

$$\text{Sigmoid}(x) = \frac{1}{1 + e^{-x}}, \tag{2.1}$$

which forces the output to be within the range of (0,1). The rectified linear unit (ReLU) is also a widely used activation function,

$$\text{ReLU}(x) = \max(0, x). \tag{2.2}$$

Finally, an activation typically used after the output layer is the softmax function. For output node i the softmax is defined as

$$\text{Softmax}(x)_i = \frac{e^{x_i}}{\sum_{j=1}^{K} e^{x_j}}, \tag{2.3}$$

where K is the total amount of output nodes. This makes sure that the output nodes are

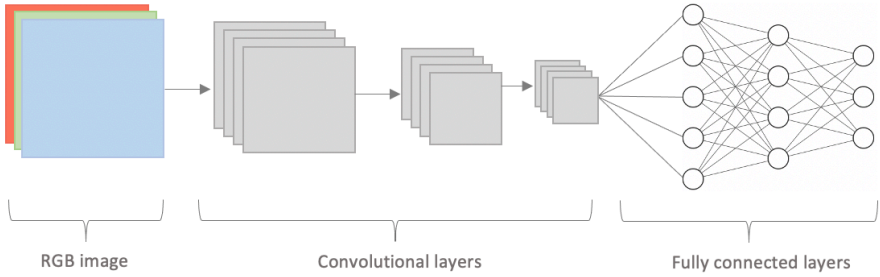

Figure 2.2: A CNN consisting of multiple convolutional layers followed by fully connected layers. The network takes an RGB image as input and outputs the prediction of three classes.

set to values within the interval (0,1), and that these values sum up to one. Hence, the softmax outputs can be interpreted as probabilities.

The neural network has the objective of minimizing a loss function. The loss function measures the error between the network's prediction and the ground truth. Two common loss functions are the mean squared error (MSE) loss and the categorical cross-entropy (CCE) loss. The MSE loss is defined as

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=0}^{N}(y_i - \hat{y}_i)^2 \,. \tag{2.4}$$

The ground truth prediction for output node i is denoted as $\hat{y}_i$ and the network's prediction as y_i. The total amount of output nodes are denoted as N. For each output node of the network a squared error will be calculated. These will be averaged in order to compute the MSE. The CCE loss is estimated by

$$\mathrm{CCE} = -\sum_{i=0}^{N}\hat{y}_i \cdot \log(y_i) \,. \tag{2.5}$$

During the training phase of a neural network the network will learn a set of parameters, i.e. weights. The training is performed by letting training data pass through the network resulting in a loss. The loss is then used for backpropagation, where partial derivatives of the loss function are derived with respect to the weights. These gradients are used in order to update the weights by following an optimizer scheme. Some common optimizers are, stochastic gradient descent (SGD) and Adam [16].

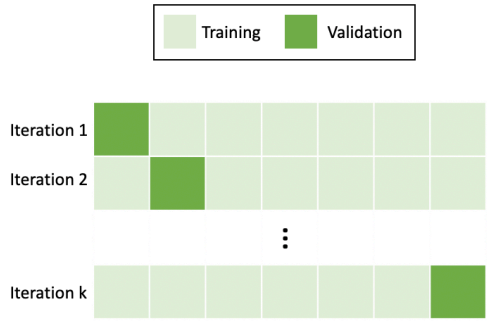

Figure 2.3: Illustration of cross-validation. For each iteration a part of the data set will be allocated for validation.

2.2 Generative Adversarial Networks

The GAN was introduced by Goodfellow et al. [8] in 2014. This is a network which can synthesise data. The key components of the GAN is a generative network G and a discriminative network D. The generative network G inputs a random latent vector and strives to generate as realistic samples as possible. The discriminative network D strives to be perfect at distinguishing the samples generated by G from real samples. The two networks D and G are trained simultaneously.

2.3 Cross-validation

When training a network it is typical to divide the data into training, validation and test sets. The training data contains samples that are used during the training phase of the network, where the network learns. To observe the learning, a validation set can be used for evaluation on unseen data. This data can be used to fine-tune the hyperparameters, and for selection of network. When the network is trained, a final evaluation is done on a test set.

In the medical field it is common to work with small data sets. Hence, splitting data into training, validation and test sets can lead to misleading conclusions. To better generalize and to obtain robust results, cross-validation can be used by splitting the data into training and validation sets multiple times. Figure 2.3 displays the principle of cross-validation, iteratively splitting the data into training and validation sets.

2.4 Metrics

There are various metrics to measure performance in ML. Important parameters are true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). These can be used to estimate the accuracy,

$$\text{Acc} = \frac{TP + TN}{TP + TN + FP + FN}, \tag{2.6}$$

the ratio between the correct predictions and all predictions. This metric is dependent on the choice of threshold for decision.

In the medical field it is paramount to measure sensitivity and specificity. Sensitivity is defined as,

$$\text{Sensitivity} = \frac{TP}{TP + FN}, \tag{2.7}$$

the probability that a patient with a disease receives an accurate diagnosis. Oppositely, the probability that a patient without the disease is correctly identified as not having the disease is defined as the specificity,

$$\text{Specificity} = \frac{TN}{TN + FP}. \tag{2.8}$$

When using accuracy in multi-class problems it can be balanced so each class has equal impact on the score. This can be expressed by using the sensitivity for each class, i.e. the probability that an entity from the class is correctly predicted. Thus, the balanced accuracy can be estimated as

$$\text{Acc}_{\text{balanced}} = \frac{1}{K} \sum_{c=0}^{K} \frac{TP_c}{TP_c + FN_c}, \tag{2.9}$$

where K denotes the amount of classes, TP_c and FN_c are the TP and FN for the class c, assuming that this class is positive and all other are negative.

The receiver operating characteristic (ROC) curve displays the performance of a classifier for each decision threshold. On the y-axis the sensitivity is shown and on the x-axis one minus the specificity is shown. From this curve it is possible to calculate the area under the ROC curve (AUC), a metric which is not dependent on a specific decision threshold,

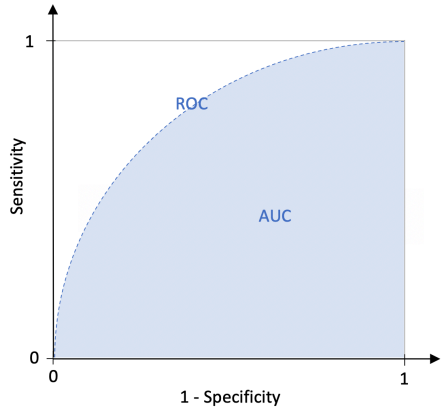

Figure 2.4: Illustration of a ROC curve, the blue area under the curve is identified as the AUC. The y-axis measures the sensitivity and the x-axis the specificity subtracted from one. Both axes are taking values in the range (0,1).

making it more robust compared to accuracy. The AUC takes a value in the range (0,1), higher values are superior. Figure 2.4 shows an illustration of a ROC curve with the AUC marked under it.

To get reliable results, a confidence interval (CI) can be used. Its lower and upper boundaries can be found by bootstrapping the test set, i.e. drawing a new test set a given number of times from the original test set. Commonly, the CI is set at 95%, which is defined as 95% confidence that the true value is within the range of the CI.

Chapter 3

Breast Ultrasound Data

DL is highly dependent on the data used during training. The data should represent the reality from which the DL should be able to make a prediction. There are several US datasets capturing breast tissue. However, the availability of POCUS data sets capturing breast is limited. This chapter will describe three different breast US data sets used in the papers covered in this work. The chapter will also include a description of important characteristics of breast lesions in US imaging, which is used by radiologist when analyzing US images.

3.1 Characteristics of Breast Lesions

Breast lesions can be classified into either benign (non-cancerous) or malignant (cancerous). Breast tissue without any lesion is referred to as normal. To distinguish between benign and malignant lesions in breast US imaging the radiologist analyzes some important characteristics of the lesion including: shape, margin, orientation, echo pattern and posterior feature of the lesion [21].

The shape of a lesion can be irregular, round or oval. Irregular structures are typically malignant and round/oval structures benign. For simplicity the margin of the lesion can be defined as circumscribed or not circumscribed, where circumscribed means that it is well-defined. For malignant lesions it is common to have a margin which is not well-defined, hence not circumscribed. The orientation of the lesion can be either parallel or non-parallel towards the surface (skin). A benign lesion is commonly parallel to the skin, whereas a non-parallel lesions could indicate a suspicious finding, possibly malignant.

The echo pattern refers to the visualization of the internal composition of the lesion. Lastly,

the posterior feature is a representation of the attenuation in the lesion due to acoustic transmission. It can be non-present or appear as shadowing or enhancement, combined or independently. Shadowing appears as a darkened area under the lesion, while enhancement will brighten the area.

3.2 Standard Breast Ultrasound Data – Egypt

Baheya hospital in Egypt has collected a publicly available breast US data set [2] consisting of images collected with standard US machine LOGIQ E9. The images are labeled to belong to one of the following three classes: normal, benign or malignant. The amount of images belonging to each class is shown in Table 3.1. This data set was used in Paper I.

Table 3.1: Total number of images for each class for the standard US data set collected at Baheya hospital, Egypt.

	Normal	Benign	Malignant	Total
US	133	437	210	780

3.3 Standard Breast Ultrasound Data – Sweden

A standard US data sets capturing breast tissue was retrospectively collected at Unilabs Mammography unit at Skåne University Hospital in Malmö, Sweden. The images were acquired with US machines LOGIQ E9 or LOGIQ E10, and labeled to one of the following three classes: normal, benign or malignant. Examples of images from each class are shown in Figure 3.1. Table 3.2 displays the amount of US images belonging to each class.

Since this data set has been expanded continuously it has been used in different versions in Paper I, II, III and IV.

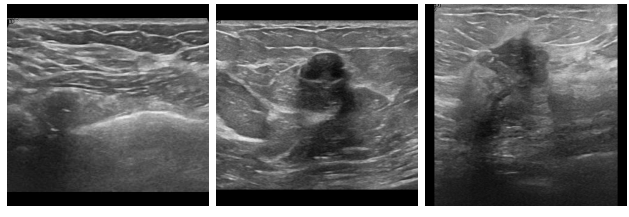

Figure 3.1: Examples of breast US images capturing normal tissue (left), benign lesion (middle) and malignant lesion (right).

Table 3.2: Total number of images of each class for the standard US data set collected at Skåne university hospital, Sweden.

	Normal	Benign	Malignant	Total
US	386	254	520	1160

3.4 Breast Point-of-Care Ultrasound Data – Sweden

The POCUS data set capturing breast tissue was collected with GE's Vscan air probe [7] at Unilabs Mammography unit at Skåne University Hospital in Malmö, Sweden. These images were assigned to either one of the classes: normal, benign or malignant. Figure 3.2 shows examples of POCUS images for each class. The data was split into a test and training set and was used in Paper II, III and IV. The same test set was used for all three papers. However, since the collection of images is an ongoing process the training sets did differ in size between the papers. Table 3.3 displays the number of POCUS images for each class and the division into test and training sets.

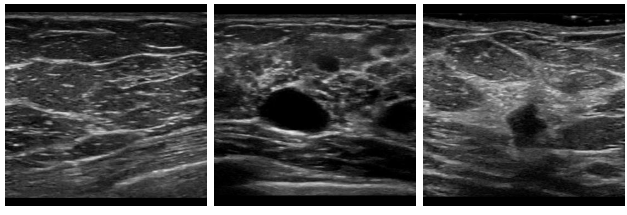

Figure 3.2: Examples of breast POCUS images capturing normal tissue (left), benign lesion (middle) and malignant lesion (right).

Images collected by different types of US devices differ in their appearance. The images of the POCUS data set consists of more pixels within a dark range compared to the images in the US data set collected in Sweden. This has been proven for data from these two data sets by investigating histograms over the pixel intensities [25].

Table 3.3: Total number of images of each class for the POCUS data set.

	Normal	Benign	Malignant	Total
Train	463	173	178	814
Test	284	131	116	531

Chapter 4

Breast Cancer Classification

In the past decade CNNs have been the state of the art for image classification tasks and have proven to perform well in various medical applications [27]. To detect breast cancer in US imaging three classes are of interest: normal, benign and malignant. In this chapter two different methods including CNNs for breast cancer classification are described: transfer learning and the use of deep features.

4.1 Transfer Learning

The principle behind transfer learning is to reuse knowledge. A neural network is trained in one setting and the gained knowledge is used in another setting. Commonly the network is trained on a large data set creating a pre-trained model. The pre-trained model is then customized to the relevant data set. Ideally the new model would have learned common structures from the pre-trained model, and specific features for the new task from the relevant data set. Figure 4.1 displays the transfer learning pipe-line.

The pre-trained model could be used in its original form with the same architecture and trained weights. However, commonly only parts of the model is retrained and the rest is frozen. The pre-trained model can also be modified by cutting away the last parts or/and adding layers suitable for the new task.

There exist several pre-trained networks which are trained on large data sets, such as ImageNet [5], and have proven to generate good results on different image classification tasks. Examples of such network architectures are ResNet, Inception, VGG and Xception [4, 11, 30, 31]. They are available in different versions where the architectures varies. In Paper I the following four pre-trained models were used: ResNet50V2, InceptionV3,

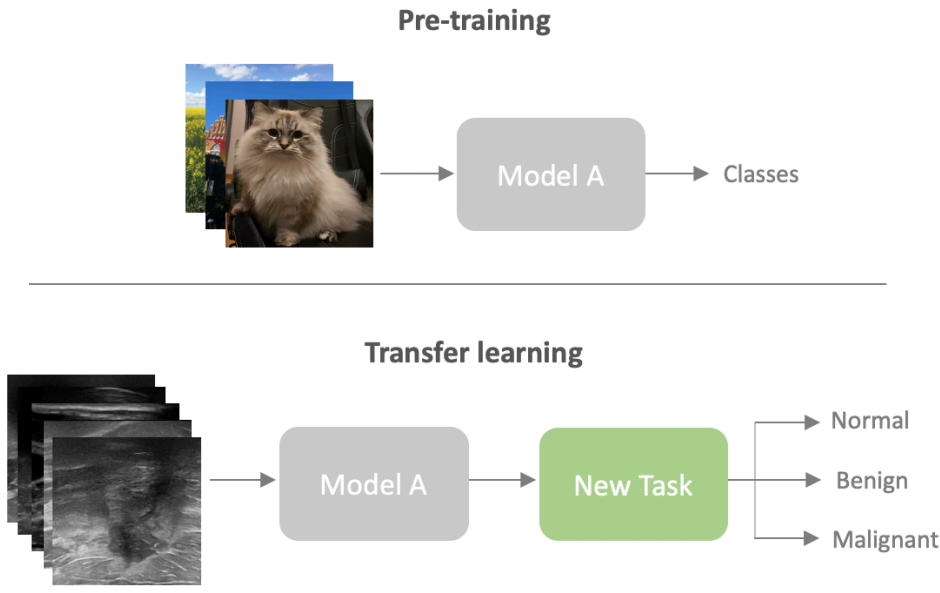

Figure 4.1: Illustration of the concept of transfer learning. Model A is pre-trained on a large data set containing multiple classes, followed by being retrained to the specific task of classifying US images.

VGG19 and Xception. To these architectures multiple fully-connected layers were added, the amount of layers and layer sizes depended on the pre-trained network.

4.2 Deep Features

Deep features refers to the features learnt by a DL model. Different DL models are possibly making different observations of the input image. Using these features combined into one large deep feature vector could enable an even better representation of the image compared to the individual features.

In Paper I large deep feature vectors were made by extracting features from different transfer learning models and concatenating them. Feature vectors can be obtained throughout different parts of the model, but it was chosen to use feature vectors from the last fully connected layer before the classification head of the network. The deep feature vectors were used as input to an ANN where the final prediction was made. Figure 4.2 illustrates the implementation of deep features in Paper I.

The results showed that using large feature vectors did not exceed the individual results of the transfer learning models. However, only 10% of the Egyptian US data set, see Section 3.2, was used for evaluation. Hence, the test set was very small making it difficult

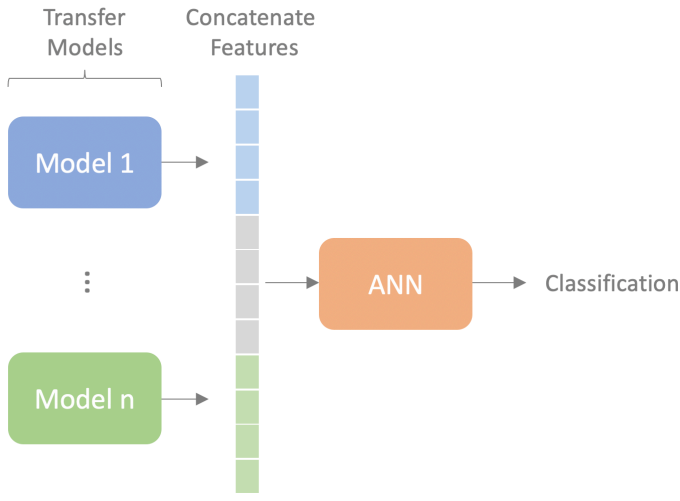

Figure 4.2: An overview of the implementation of deep features in Paper I. The obtained feature vectors from n transfer learning models are concatenated into one deep feature vector used as input into an ANN making the final classification.

to draw proper conclusions.

Chapter 5

Managing Scarcity of Data

A common issue working with medical data is the difficulty of collecting it. Medical data is capturing sensitive personal information and requires approvals and consents, which can be a time-consuming process. There are different ways of increasing a data set without the collection of new data. The data set could be expanded by using augmentation or by generating new data using GANs. In this chapter two methods for increasing data sets will be covered. Starting with description of augmentation, followed by domain shift of images using a cycle-consistent adversarial network (CycleGAN).

5.1 Augmentation of Breast Ultrasound Images

Data augmentation is a method where the training images are copied and altered in order to increase the variability within the data. There are different approaches to augmentation, such as, spatial altering, color shift, changing brightness and adding noise.

Since US images are in grayscale, color shift is not of high importance. However, since images acquired with different US machines have different distributions of pixel intensities and contain different amounts of noise, these might be of interest to alter in order to broaden the content of the data set.

Spatial augmentation has proven to be useful for increasing the classification performance of breast US images [1]. When applying spatial augmentation to US images it is important to alter with moderation. This to prevent the characteristics mentioned in Section 3.1 to be damaged. For example, a too heavy rotation could compromise with the orientation of the lesion which is an important characteristic in distinguishing cancer versus non-cancer.

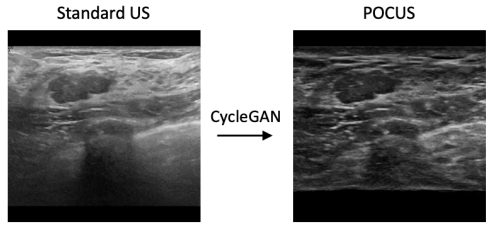

Figure 5.1: A standard US images (left) is shifted into the domain of POCUS images (right) by using CycleGAN.

5.2 Domain Shift Using Cycle-Consistent Adversarial Networks

As previously mentioned the access to breast POCUS data is limited and collecting new data in the clinics takes a lot of resources and is time consuming. Generating unseen images from scratch without physically collecting them, for example by using GANs, is risky. Particularly for a field such as medical diagnostics, as important structures risk being overlooked. However, the availability of standard breast US data is greater since this is a common examination method. Thus, these images can be retrieved retrospectively. CycleGANs are networks which can be trained to shift images from one domain into another domain [36]. For example, it could shift an image of a horse into an image of a zebra and vice versa. As the access to standard breast US images is superior to POCUS images, the POCUS data set could be expanded by shifting the standard US images into the domain of POCUS. Since the CycleGAN shifts the image domain, important structures should be preserved and not overlooked. Thus, making it potentially more safe compared to generating unseen images with GANs. Figure 5.1 displays how a standard US image is shifted into the domain of POCUS images.

CycleGANs are built upon the idea of GANs. The CycleGAN consists of two image domains X and Y. Each domain has its own generator G, and discriminator D. The generator for domain X is denoted as G_X and it takes an images from X as input, striving to translate this image to appear as being from domain Y. In the same way there is a generator G_Y which translates images from domain Y to appear as being from domain X. In addition to the generators there are two discriminators, one for each domain, denoted as D_X and D_Y. These are striving to be as good as possible in distinguishing the translated images from real images. The CycleGAN should be cycle-consistent, which means that if an image is translated two times it should go back to its original appearance. Thus, the result of inputting image x from domain X into the generators, $G_Y(G_X(x))$, should be as close to image x as possible. The same should be true for inputting an image y from domain Y into $G_X(G_Y(y))$. The CycleGAN method was used in Paper II, showing that the classification performance improved when adding the CycleGAN generated POCUS images to the POCUS training set.

Chapter 6

Towards a Trustworthier Classifier

Inaccurate diagnosis can lead to serious implications. Hence, it is crucial to employ safety mechanisms to assure trustworthy results of a DL classifier. An important step is the ability for the classification network to detect OOD samples. These are samples outside the networks knowledge, i.e. too far away from the in-distribution (ID) data. Furthermore, the classification network should be able to tell when the prediction of a sample from the ID data is not trustworthy. Thus UQ is important, i.e. to estimate how secure the network is in its prediction.

There are many methods for both OOD detection [35] and UQ [23]. In this chapter the main concepts of OOD detection and UQ will be described. These are followed by a section covering some OOD detection and UQ methods: softmax score, energy score, deep ensemble, and Bayesian neural networks (BNNs).

6.1 Out-of-Distribution Detection

Several different types of OOD samples exist in breast US imaging. These include images capturing non-breast tissue, images of poor quality and images capturing rare lesions. A network specified in classifying breast tissue should not be used on structures such as bone or artery, since these have not been used in neither training or evaluation of the classifier. Hence, other structures than breast tissue should be detected as OOD samples and no prediction should be made. Furthermore, US imaging is difficult. Using an inadequate amount of ultrasonic gel or not applying enough pressure with the probe will lead to images of poor quality. Images of bad quality due to faulty managing of the US probe should be discarded since important characteristics could be compromised. The safety mechanism of OOD detection could be paramount when a non-experienced examiner uses the DL

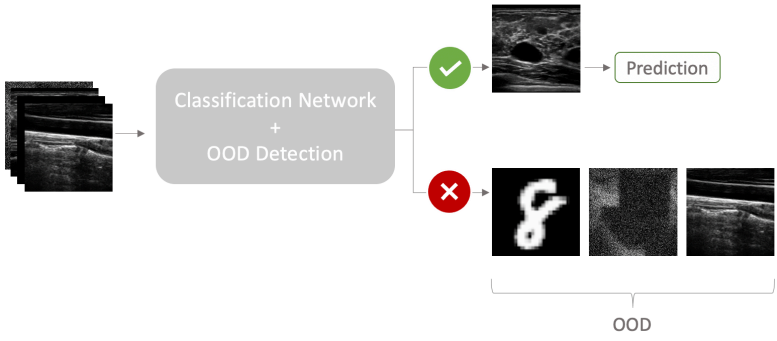

Figure 6.1: Overview of a classification network including OOD detection. Images which are OOD should not be used for prediction.

classifier. Figure 6.1 displays the concept of a classification network combined with OOD detection.

The OOD detection is done by calculating a score for the sample and comparing it to a set threshold for distinguishing OOD samples from ID samples. There are different methods for OOD detection such as, softmax score, energy score and deep ensembles. These are introduced in Sections 6.3.1, 6.3.2 and 6.3.3 respectively.

6.2 Uncertainty Quantification

UQ is a way of measuring the uncertainty of the classification network's output. A high uncertainty should be connected to a poor classification result and a low uncertainty to a trustworthy result. Hence, UQ is useful to find samples for which the predictions should not be trusted, for example samples including rare lesions. When the uncertainty is high, the first step will be to acquire a new images. If the uncertainty is not decreasing the sample can not be predicted properly by the classification network. An overview of the pipe-line for applying UQ to the classification network is shown in Figure 6.2.

The uncertainty can be measured as the total uncertainty (TU), which consists of an aleatoric and an epistemic part [12]. The aleatoric uncertainty (AU) is related to the nature of the data in terms of noise and variability. Examples of AU is measurement errors or simply that the image cannot be explained by the label. Thus, adding more data would not decrease the AU since it is irreducible by nature [17]. Epistemic uncertainty (EU) is defined as the uncertainty within the model itself, referring to its lack of knowledge about underlying data distributions or class boundaries. Hence, adding more data and increasing the variability within the data should decrease the EU [17]. To calculate the TU, AU and EU an uncertainty metric is needed, denoted as H. The TU is defined as

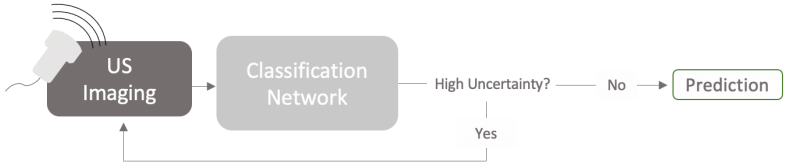

Figure 6.2: Overview of UQ applied to the classification network. A high uncertainty implies that a new image should be collected in order to make a safe prediction.

$$\mathcal{U}_{tot} = H[\mathbb{P}(y|x, D)]\,, \tag{6.1}$$

where $\mathbb{P}(y|x, D)$ is the predictive distribution, given input image x, and the observed data D. The parameters of the model, θ, are assumed to be realisations of the stochastic variable, Θ. The AU is defined as

$$\mathcal{U}_{ale} = \mathbb{E}_{\Theta|D}[H[\mathbb{P}(y|x, \theta)]]\,. \tag{6.2}$$

Finally the EU can be obtained as following,

$$\mathcal{U}_{epi} = \mathcal{U}_{tot} - \mathcal{U}_{ale}\,. \tag{6.3}$$

The uncertainty metric H can be defined in multiple ways. For classification tasks the most common uncertainty metric is entropy [26]. In general the entropy is defined as

$$H_{entropy}(p) = H_{entropy}(p_1, p_2, ..., p_n) = -\sum_{i=1}^{n} p_i \cdot \log p_i\,, \tag{6.4}$$

where p denotes a vector of probabilities of length n. Another common uncertainty metric is to use the variance, which is widely used in regression tasks [26]. Generally the variance for a vector p of length n is defined as

$$H_{variance}(p) = \mathrm{Var}[p_1, p_2, ..., p_n] = \frac{1}{n}\sum_{i=1}^{n}(p_i - \bar{p})^2\,, \tag{6.5}$$

where $\bar{p}$ is the mean.

UQ could be implemented for OOD detection, setting a threshold for the uncertainty score distinguishing OOD samples from ID samples.

6.3 OOD Detection and UQ Methods

6.3.1 Softmax Score

The softmax score is obtained as probabilities from the softmax activation function of the last layer in a network. A low score should identify the predicted sample as OOD. For UQ a high softmax score should imply a low uncertainty. A problem with the softmax score is that the probabilities for the classes are required to sum up to one which can be problematic if none of the classes fits the sample. It has previously been shown that the softmax score is not reliable for neither UQ or OOD detection [24].

6.3.2 Multilevel Energy Score

Energy score is a method where the logits of a network are used to decide whether a sample is ID or OOD [20]. Logits are collected before the softmax activation function, these are the unscaled scores of the network. Unlike softmax, they do not need to sum up to one. The energy score of sample x and network f is defined as

$$E(x;f) = -T \cdot \log \sum_{c=1}^{K} e^{f_c(x)/T}, \tag{6.6}$$

where $f_c(x)$ denotes the logit for class c, the total amount of classes is written as K, and the score is scaled by a temperature parameter T.

It has been suggested to investigate logits collected from different parts throughout the network. This is called multi-level out-of-distribution detection (MOOD) and was proposed by Lin et al. [19]. The idea is that simple OOD samples would be easily detected at an early stage of feature extraction in the network, while more complex samples would be detected further on into the network.

The energy score using multiple exits inspired by MOOD was implemented in Paper III. In the paper it was shown that different types of OOD samples were detected at different stages of the network, showing the value of using multiple exits for OOD detection with energy score.

6.3.3 Deep Ensemble

The principle behind deep ensembles is to train multiple networks independently and combine their outputs in order to make a more informed and better prediction [10]. Ensembles

could increase reliability and generalization of the network [22]. The combined outputs could also be used to obtain uncertainties. The uncertainties are obtained by UQ and could further be used for OOD detection.

There are various ways of combining several neural networks into a deep ensemble. Each individual network of the deep ensemble is called an ensemble member. One way of combining the predictions of the ensemble members is to use the average of their predictions. The output of an average ensemble for class c with N members, is defined as

$$\overline{M}_c(x) = \frac{1}{N}\sum_{n=1}^{N} M_c^{(n)}(x)\,, \tag{6.7}$$

where the prediction of ensemble member n is denoted as $M_c^{(n)}$, and x is the input sample. Another method is to let the ensemble members vote. Hence, the prediction for the final class would be set as the class with majority amongst the members.

The ensemble members can be used to calculate the uncertainty. As mentioned in Section 6.2 there are several uncertainties that could be derived, see Equations 6.1, 6.2 and 6.3. Selecting variance-based uncertainty as uncertainty metric, the TU of the ensemble can be calculated as

$$\mathcal{U}_{variance}^{tot} = \sum_{c=1}^{K}\frac{1}{N}\sum_{n=1}^{N}(M_c^{(n)}(x) - \overline{M}_c(x))^2\,. \tag{6.8}$$

The variance-based TU can be used with weights in order to give the predicted class more impact [34],

$$\mathcal{U}_{weighted}^{tot} = \sum_{c=1}^{K}\overline{M}_c(x)\frac{1}{N}\sum_{n=1}^{N}(M_c^{(n)}(x) - \overline{M}_c(x))^2\,. \tag{6.9}$$

Using entropy as uncertainty metric, the TU of the ensemble is defined as

$$\mathcal{U}_{entropy}^{tot} = -\sum_{c=1}^{K}\overline{M}_c(x)\log(\overline{M}_c(x))\,. \tag{6.10}$$

6.3.4 Bayesian Neural Networks

As opposed to regular neural networks in which the weights are fixed values, the weights of a BNN are probability distributions over a range of possible values [18]. The BNN seeks to

learn the posterior distribution, i.e. the probability distribution of weights w, given training data D_{train}. This is computed with Bayes rule,

$$p(w|D_{train}) = \frac{p(D_{train}|w)p(w)}{p(D_{train})} . \tag{6.11}$$

To calculate the posterior distribution this way requires both the prior $p(w)$ and the likelihood $p(D_{train}|w)$ to be specified. Thus, all possible weights needs to be considered, making infinitely many options to evaluate. Hence, approximation can be used. There are different methods to find such approximations. Monte Carlo dropout can be used to simulate a BNN and is fairly simple to implement [6]. For a fixed amount of times inference will be performed with random dropout. Dropout will generate different predictions which can be used to compute the uncertainty of the prediction in an ensemble-like style.

Chapter 7

Discussion

This thesis covers the topic of classification of breast cancer in US imaging. The goal is to develop a support tool for breast cancer detection suitable in LMICs, which could be used by a non-experienced examiner, i.e. non-radiologist. In order to make a useful support tool there are multiple important parts that need to be in place. The classification performance should be at least at the level of a radiologist and preferably independent of the brand of the US probe. Paper II shows that using data augmentation during training of the classification network plays an important part in improving the performance. However, even though good results could be achieved, the support tool should be able to detect OOD samples and warn the examiner when a prediction should not be made. In Paper III such a mechanism was investigated when different OOD detection methods were evaluated. It was shown that the different methods had different strengths for detecting the OOD samples. For future work, more methods should be investigated. One should also consider combining different methods. Except OOD detection, UQ is important in order to implement a trustworthy classifier, aiming to have low uncertainty correlated with good performance. Paper IV compares three ways of calculating uncertainty scores, showing that the performance of the classifier improves when samples with high uncertainty are excluded. This implies that these samples require further examination.

For future work there are a lot of areas to investigate and challenges that needs to be solved. One challenge is that as of today there are several POCUS probes of different brands. The ultimate goal is to have a classification network that works well independent of the brand of the probe. Hence, future work should include methods for increasing the generalizability of the classification network. As mentioned in Section 3.4 the quality and appearance of images collected with different US machines varies. Transfer learning could enable the classification network to make correct predictions independently of the US machine used to collect the data. This could be implemented by customizing a pre-trained model trained

on standard US to classifying POCUS images acquired with probes of different brands.

Further, one challenge in US imaging is the managing of the probe, especially for non-experienced examiners. ML has the potential of guiding the examiner during the examination. The ML guide could for example notify the examiner if more ultrasonic gel is needed, if the angle of the probe should be altered or if the probe should be used with more pressure. Adding such a guiding system to the support tool could make the examination faster and more accurate. The classification networks trained in the mentioned papers are trained on individual images. However, US imaging is often acquired by short video sequences. Training the classifier on these could be of interest since they contain more information than one individual image.

For the proposed support tool to be useful, clinical studies needs to be performed to show the feasibility in real world settings. The data used in this work is mainly collected in Sweden. Thus, there is a risk of it being biased and possibly not representable to LMICs. Hence, this needs to be further studied. There are many challenges to be solved in order for the proposed support tool to be useful in LMICs. However, this work shows that there is potential for a DL classifier to detect breast cancer in POCUS images in a trustworthy way.

References

[1] W. Al-Dhabyani, M. Gomaa, H. Khaled, and A. Fahmy. Deep learning approaches for data augmentation and classification of breast masses using ultrasound images. *International Journal of Advanced Computer Science and Applications*, 10(5), 2019. doi: 10.14569/IJACSA.2019.0100579. URL http://dx.doi.org/10.14569/IJACSA.2019.0100579.

[2] W. Al-Dhabyani, M. Gomaa, H. Khaled, and A. Fahmy. Dataset of breast ultrasound images. *Data in Brief*, 28:104863, 2020. ISSN 2352-3409. doi: https://doi.org/10.1016/j.dib.2019.104863. URL https://www.sciencedirect.com/science/article/pii/S2352340919312181.

[3] F. Bray, M. Laversanne, H. Sung, J. Ferlay, R. L. Siegel, I. Soerjomataram, and A. Jemal. Global cancer statistics 2022: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 74(3):229–263, 2024. doi: https://doi.org/10.3322/caac.21834. URL https://acsjournals.onlinelibrary.wiley.com/doi/abs/10.3322/caac.21834.

[4] F. Chollet. Xception: Deep learning with depthwise separable convolutions, 2017. URL https://arxiv.org/abs/1610.02357.

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

[6] Y. Gal and Z. Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning, 2016. URL https://arxiv.org/abs/1506.02142.

[7] GE Healthcare. Vscan air. https://vscan.rocks/product/vscanair Accessed: 2024-08-12.

[8] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*, 2014. URL https://arxiv.org/abs/1406.2661.

[9] F. Guida, R. Kidman, J. Ferlay, J. Schüz, I. Soerjomataram, B. Kithaka, O. Ginsburg, R. B. Mailhot Vega, M. Galukande, G. Parham, S. Vaccarella, K. Canfell, A. M. Ilbawi, B. O. Anderson, F. Bray, I. dos Santos-Silva, and V. McCormack. Global and regional estimates of orphans attributed to maternal cancer mortality in 2020. *Nature Medicine*, 28(12):2563–2572, 2022. doi: 10.1038/s41591-022-02109-2. URL https://doi.org/10.1038/s41591-022-02109-2.

[10] L. Hansen and P. Salamon. Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(10):993–1001, 1990. doi: 10.1109/34.58871.

[11] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition, 2015. URL https://arxiv.org/abs/1512.03385.

[12] E. Hüllermeier and W. Waegeman. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Machine Learning*, 110(3):457–506, 2021. doi: 10.1007/s10994-021-05946-3. URL https://doi.org/10.1007/s10994-021-05946-3.

[13] J. Karlsson, J. Ramkull, I. Arvidsson, A. Heyden, K. Åström, N. C. Overgaard, and K. L. M.D. Machine learning algorithm for classification of breast ultrasound images. In K. Drukker and K. M. Iftekharuddin, editors, *Medical Imaging 2022: Computer-Aided Diagnosis*, volume 12033, page 120331T. International Society for Optics and Photonics, SPIE, 2022.

[14] J. Karlsson, I. Arvidsson, F. Sahlin, K. Åström, N. C. Overgaard, K. Lång, and A. Heyden. Classification of point-of-care ultrasound in breast imaging using deep learning. In K. M. Iftekharuddin and W. Chen, editors, *Medical Imaging 2023: Computer-Aided Diagnosis*, volume 12465, page 124650Y. International Society for Optics and Photonics, SPIE, 2023.

[15] J. Karlsson, M. Wodrich, N. C. Overgaard, F. Sahlin, K. Lång, A. Heyden, and I. Arvidsson. Towards out-of-distribution detection for breast cancer classification in point-of-care ultrasound imaging. In *Lecture Notes in Computer Science*. 27th International Conference on Pattern Recognition, Springer Nature, 2024.

[16] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization, 2017. URL https://arxiv.org/abs/1412.6980.

[17] A. D. Kiureghian and O. Ditlevsen. Aleatory or epistemic? does it matter? *Structural Safety*, 31(2):105–112, 2009. ISSN 0167-4730. doi: https://doi.org/10.1016/

j.strusafe.2008.06.020. URL https://www.sciencedirect.com/science/article/pii/S0167473008000556. Risk Acceptance and Risk Communication.

[18] J. Lampinen and A. Vehtari. Bayesian approach for neural networks—review and case studies. *Neural Networks*, 14(3):257–274, 2001. ISSN 0893-6080. doi: https://doi.org/10.1016/S0893-6080(00)00098-8. URL https://www.sciencedirect.com/science/article/pii/S0893608000000988.

[19] Z. Lin, S. D. Roy, and Y. Li. Mood: Multi-level out-of-distribution detection, 2021.

[20] W. Liu, X. Wang, J. D. Owens, and Y. Li. Energy-based out-of-distribution detection, 2021. URL https://arxiv.org/abs/2010.03759.

[21] E. Mendelson, M. Böhm-Vélez, and a. Berg. *ACR BI-RADS® Ultrasound. In: ACR BI-RADS® Atlas, Breast Imaging Reporting and Data System.* American College of Radiology, 2013.

[22] A. Mohammed and R. Kora. A comprehensive review on ensemble deep learning: Opportunities and challenges. *Journal of King Saud University - Computer and Information Sciences*, 35(2):757–774, 2023. ISSN 1319-1578. doi: https://doi.org/10.1016/j.jksuci.2023.01.014. URL https://www.sciencedirect.com/science/article/pii/S1319157823000228.

[23] V. Nemani, L. Biggio, X. Huan, Z. Hu, O. Fink, A. Tran, Y. Wang, X. Zhang, and C. Hu. Uncertainty quantification in machine learning for engineering design and health prognostics: A tutorial. *Mechanical Systems and Signal Processing*, 205:110796, 2023. ISSN 0888-3270. doi: 10.1016/j.ymssp.2023.110796. URL http://dx.doi.org/10.1016/j.ymssp.2023.110796.

[24] A. M. Nguyen, J. Yosinski, and J. Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. *CoRR*, abs/1412.1897, 2014. URL http://arxiv.org/abs/1412.1897.

[25] F. Sahlin. Detection of breast cancer in pocket ultrasound images using deep learning. *Lund University: Master's Theses in Mathematical Sciences 2022:E16*, pages ISSN: 1404–6342, 2022. ISSN 1404-6342.

[26] Y. Sale, P. Hofman, L. Wimmer, E. Hüllermeier, and T. Nagler. Second-order uncertainty quantification: Variance-based measures, 2023. URL https://arxiv.org/abs/2401.00276.

[27] D. R. Sarvamangala and R. V. Kulkarni. Convolutional neural networks in medical image understanding: a survey. *Evolutionary Intelligence*, 15(1):1–22, 2022.

[28] S. Shamshirband, M. Fathi, A. Dehzangi, A. T. Chronopoulos, and H. Alinejad-Rokny. A review on deep learning approaches in healthcare systems: Taxonomies, challenges, and open issues. *Journal of Biomedical Informatics*, 113:103627, 2021. ISSN 1532-0464. doi: https://doi.org/10.1016/j.jbi.2020.103627. URL https://www.sciencedirect.com/science/article/pii/S1532046420302550.

[29] R. L. Siegel, K. D. Miller, H. E. Fuchs, and A. Jemal. Cancer statistics, 2022. *CA: A Cancer Journal for Clinicians*, 72(1):7–33, 2022.

[30] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition, 2015. URL https://arxiv.org/abs/1409.1556.

[31] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions, 2014. URL https://arxiv.org/abs/1409.4842.

[32] A. Tfayli, S. Temraz, R. Abou Mrad, and A. Shamseddine. Breast cancer in low- and middle-income countries: An emerging and challenging epidemic. *Journal of Oncology*, 2010(1):490631, 2010.

[33] WHO. The global breast cancer initiative, 2022. https://www.who.int/publications/m/item/the-global-breast-cancer-initiative-gbci Accessed: 2024-08-12.

[34] M. Wodrich, J. Karlsson, K. Lång, and I. Arvidsson. Trustworthiness for deep learning based breast cancer detection using point-of-care ultrasound imaging in low-resource settings. In *Communications in Computer and Information Science*. MICCAI meets Africa Worskshop at the 27th International Conference on Medical Image Computing and Computer Assisted, Springer Nature, 2024.

[35] J. Yang, K. Zhou, Y. Li, and Z. Liu. Generalized out-of-distribution detection: A survey, 2024. URL https://arxiv.org/abs/2110.11334.

[36] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017.

Scientific Publications

Printed by Media-Tryck, Lund 2024 NORDIC SWAN ECOLABEL 3041 0903

Licentiate Thesis in Mathematical Sciences 2024:2

ISBN 978-91-8104-157-6
ISSN 1404-028X