UNIVERSITY OF OULU

FACULTY OF MEDICINE

Mustafa Al-Rubaye

AUTOMATIC QUALITY ASSESSMENT IN MAMMOGRAPHY SCREENING: A DEEP LEARNING BASED SEGMENTATION METHOD

Bachelor’s Thesis
Degree Programme in Medical and Wellness Technology
May 2021

Al-Rubaye M. (2021) Automatic Quality Assessment in Mammography Screening: A Deep Learning Based Segmentation Method. University of Oulu, Degree Programme in Medical and Wellness Technology, 34 p.

ABSTRACT

Mammography is an imaging method used as a main tool to detect breast cancer at early stages. Images (mammograms) are examined by radiologists, who aim to identify cancerous findings. However, in order to do that, the mammograms need to be of diagnostic quality, which can sometimes be insufficient, and thus the quality of diagnosis also suffers.

Radiology technicians (radiographers) are trained to take mammography images, but not in every healthcare center a strict quality control process is established, which may substantially affect the patients. The most common defects in mammograms are positioning defects, which are seen in the images as skin-foldings or non-imaged parts of the breast.

The major issue at a process level is that the described positioning issues are noticed late, already at the diagnostic phase. If a radiologist decides that the mammogram is a non-diagnostic quality, the patient needs to revisit the imaging center. If quality control could be automated and standardized, unnecessary patient recalls could be avoided, thus, reducing the costs of the mammographic process. To date, there is a lack of automatic general quality control tools for mammography screening. Looking at the recent advances in artificial intelligence, it may be possible to automate this process.

The goal of this thesis was to develop an automatic system for quality assessment of mammograms. The author used Deep learning to develop an automatic framework for automatic segmentation of defects in mammograms using a dataset of 512 mammographic images extracted from the Oulu University Hospital archive. The second stage of the developed method performed quality assessment by analyzing the presence and location of different tissues in the images from the predicted segmentations.

The developed segmentation model yielded a Dice coefficient over 0.90 for the whole breast, breast, and pectoral muscle, and over 0.60 for skin-foldings and nipple.

The developed method is the first to tackle automatic segmentation of all major positioning issues in mammography. Ultimately, the developed technology has a potential to improve the mammography workflows and, eventually, patient outcomes.

Keywords: mammography, quality assessment, proof-of-concept, deep learning, segmentation

Al-Rubaye M. (2021) Automaattinen laadunarviointi mammografian kuvauksessa: syväoppimispohjainen segmentointimenetelmä. Oulun yliopisto, Hyvinvointitekniikan koulutusohjelma, 34 s.

TIIVISTELMÄ

Mammografiaa on kuvantamismenetelmä, jota käytetään päävälineenä rintasyövän havaitsemiseksi varhaisessa vaiheessa. Radiologien on tutkittava mammogrammit ja päätettävä sitten, onko pahanlaatusia löydöksiä, ja tätä varten mammografiakuvien on oltava diagnostisesti laadukkaita.

Ammattilaiset koulutetaan mammografiakuvien ottamiseksi, mutta ei kaikissa terveyskeskuksissa on otettu käyttöön tiukka laadunvalvontaprosessi, joka voi vaikuttaa merkittävästi potilaisiin. Kuvissa voi olla virheitä, jotka tekevät kuvista ei-diagnostisen laadukkaan mammogrammin, ja ne voivat vaikuttaa diagnostiikkatuloksiin. Yksi näistä vioista ovat paikannusvirheet, joissa näkyvät kuvissa ihon taitoksina ja jotkut rinnan osat eivät näy.

Suurin ongelma prosessitasolla on, että kuvatut paikannusvirheet havaitaan myöhässä, jo diagnoosivaiheessa. Jos radiologit päättävät, että mammografiakuva ei ole diagnostisesti laadukas, potilaan on palattava kuvantamiskeskukseen ja tutkittava uudelleen, mikä voi lisätä kustannuksia ja työmäärää. Jos laadunvalvonta voidaan automatisoida ja standardoida, voidaan välttää tarpeetonta potilaan palauttamista ja vähentää siten mammografiaprosessin kustannuksia. Tähän mennessä mammografiaseulonnassa ei ole automaattista yleistä laadunvalvontaa. Kun tarkastellaan tekoälyn viimeaikaisia edistystä, tämän prosessin automatisointi voi olla mahdollista.

Tämän projektin tarkoituksena oli todistaa diagnostisten ja ei-diagnostisten laatumammogrammien automaattisen erottamisen toteutettavuus. Kirjoittaja käytti syvää oppimista automatisoidun kehyksen luomisessa käyttämällä 512 mammografiakuvaa, jotka otettiin Oulun yliopistollisen sairaalan arkistosta. Automaattisen menetelmän ensimmäisessä vaiheessa suoritettiin rintakudosten ja ihon taittumien segmentointi. Toisessa vaiheessa suoritettiin laadunarviointi analysoimalla eri kudosten läsnäolo ja sijainti kuvissa.

Kehitetyllä segmentointimallilla saavutettiin merkittäviä tuloksia, kun koko rinnan ja rintalihasten segmentoinnin onnistumisen hyvyttä mittaava Dice-kerroin oli yli 0,90, ja ihon taittumiselle ja nännille yli 0,60.

Kehitetty menetelmä on ensimmäinen, joka käsittelee mammografian kaikkien tärkeimpien paikannusvirheiden automaattista segmentointia. Sillä on potentiaalia myötävaikuttaa mammografian työnkulkujen ja potilastulosten parantamiseen.

Avainsanat: mammografia, laadunarviointi, käsitteen todistaminen, syväoppiminen, segmentointi

TABLE OF CONTENTS

1. INTRODUCTION

Breast cancer (BC) is the most common type of cancer in females in the world [1, 2]. Mammography is one of the imaging methods used to detect BC. Technically, it is a process of using low-energy X-rays (around 30 kVp) to examine human breast, and it is considered to be the most reliable screening tool for early BC prevention [3, 4].

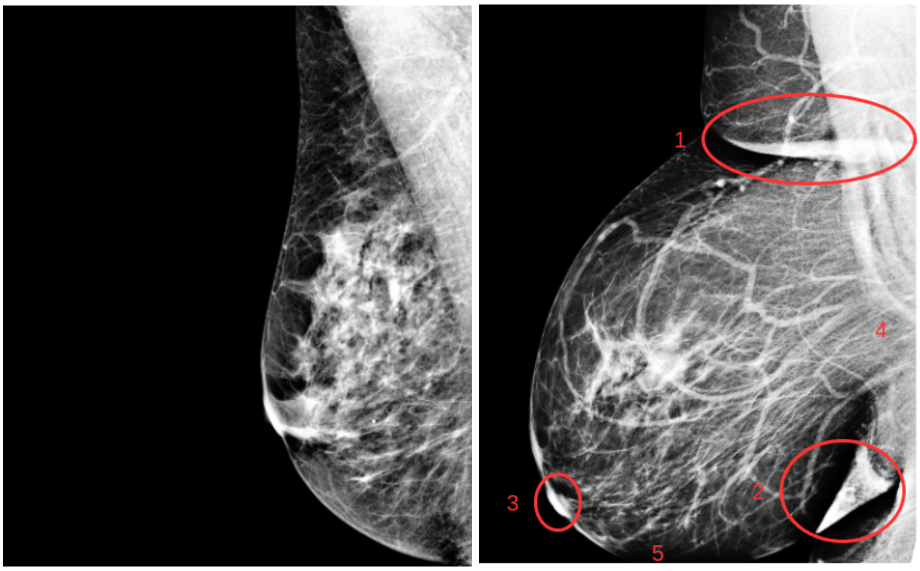
During a mammography exam, a radiographer (technician) will perform the positioning of a patient and the imaging device, with a subsequent compression of the patient breast between two plates. The compression is necessary to spread the breast tissue and to eliminate motion, which can blur the image [5]. An X-ray image of the breast is then taken in such a state [5]. Images will be stored digitally in a computer, as nowadays most of the mammograms are digital. Dense breast tissue, like cancer, appears brighter and whiter than less dense tissues, such as fat that appears dark or in gray. After the imaging is done, a radiologist (diagnosing doctor) will examine the image and give a diagnosis. Importantly, prior to the image presented to the radiologist, it undergoes a quality check, done by a radiographer [6] in the very beginning of the data acquisition.

Sustaining implementation of quality in images is among the important challenges in mammography [7]. Proper positioning is needed to perform mammography screening, since incorrect positioning and imaging artifacts often result in non-diagnostic quality images (Figure 1). The radiologist (doctor) is responsible only for the clinical aspects of mammography quality (i.e. reliable diagnosis) and he/she relies on the compliance of a radiographer with the quality standards [7].

Positioning of the breast is one of the challenges related to technical performance of mammographic process [8], and it is considered to be one of the key factors affecting the diagnostic outcomes of mammography [9]. Mispositioning may reduce the quality of the images and also hide potential abnormalities behind skin folds, thereby causing interpretation errors [6]. Typically, if the image is suspected to be non-diagnostic, a patient is asked to visit the imaging center again (process known as recall) [10], which causes additional costs and additional inconvenience to the patient.

If the process of assessing mammography imaging quality could be automated, it might substantially help with the aforementioned issues and save costs by saving radiographers' time, reduce the number of recalls, and also shorten the diagnostic time.

Positioning of the breast during the acquisition of a mammogram can be assessed by analyzing multiple factors, such as length and shapes of different parts of the breast in the image. To detect and localize different tissues and abnormalities within the image, automatic segmentation tools can be applied [11]. In this project, the author of the present thesis developed a novel method to automatically segment breast tissues and imaging defects in mammography images. The core of this method is based on Deep Learning, which is primarily discussed in the present work.



(a) (b)

Figure 1. Examples of diagnostic and non-diagnostic images – (a) and (b), respectively. In subfigure (b), skin-foldings are present (denoted by 1 and 2), nipple (denoted by 3) is too low and is not on the same line as the pectoral muscle (denoted by 4). Finally, some parts of the breast tissue are not fully visible in the mammogram (denoted by 5).

2. BACKGROUND

2.1. Breast Cancer

Disease. Breast cancer (BC) is a cancer that develops in the breast cells [12]. BC can occur in men and women and may have different phenotypes even though breast carcinomas have some similar characteristics in both genders [13]. In general, there is not a lot of research on men with BC, and most knowledge on male BC are obtained from studies on females [13, 14], as it is the most common type of cancer in the world [1, 2].

Female breast is the tissue covering the pectoral muscle. The breast is made of glandular tissue that produces milk and fatty tissue. The milk is produced in the lobules, that is located in the lobes. Milk then travels through the ducts, which is a network of tiny tubes. The tubes come together and make larger ducts that exit the skin eventually in the nipple. The ducts tubes carry the milk to the nipple. The nipple is surrounded with dark area of skin called the areola. Breast shape comes from different ligaments and connective tissue that provide support (Figure 2). The breast also contains nerves that provide sensation, blood and lymph vessels, and lymph nodes [15].

Cancer can start to manifest itself in different parts of the breast (Figure 3). Most BCs in women take place in the ducts or lobules [16, 17], and can then spread to other parts of the body through blood and lymph vessels [18].

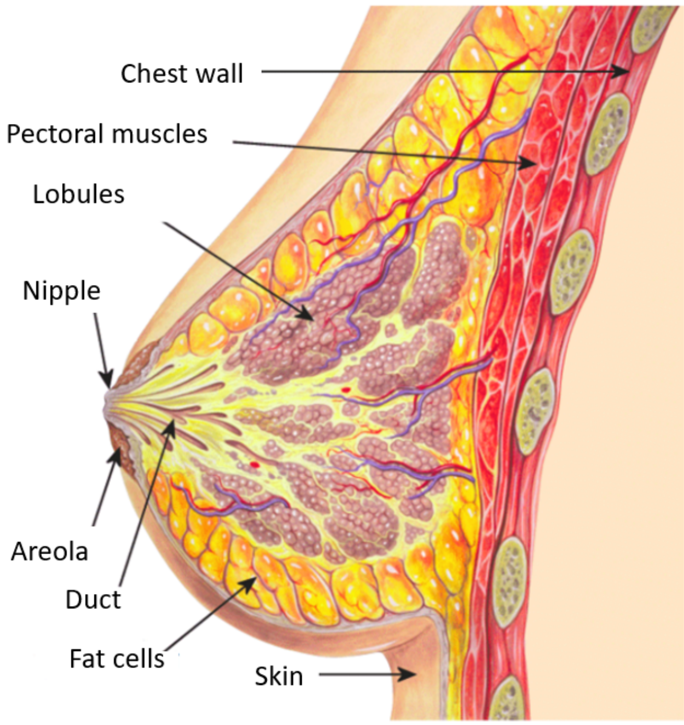
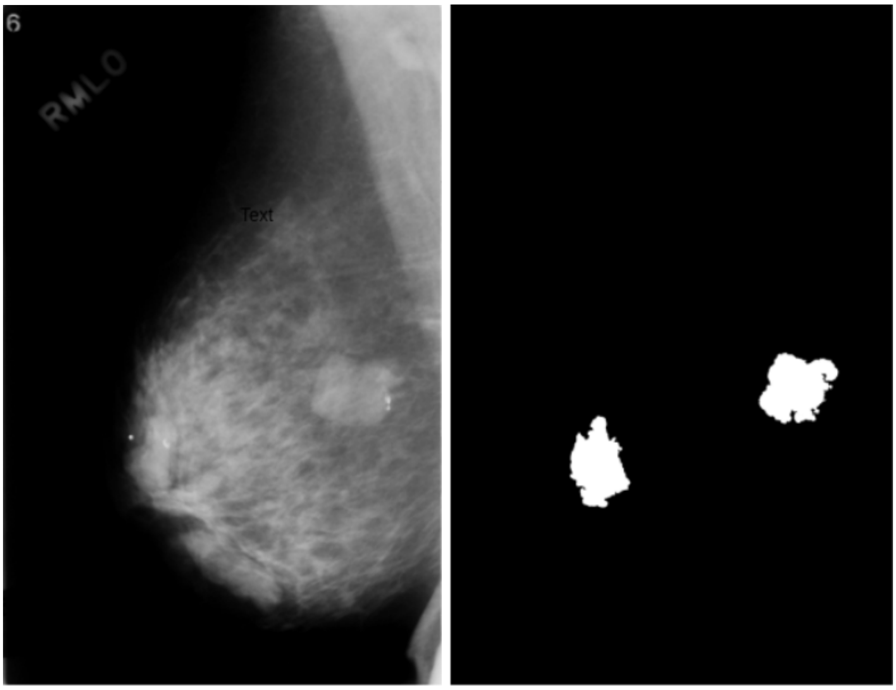


Figure 2. Breast normal anatomy cross-section [19].

Economical burden. In the United States, 255,180 new BC cases and 41,070 deaths from BC were estimated in 2017 [20]. The annual cost of medicare for BC



(a) (b)

Figure 3. An example of mammography image (a) and the corresponding cancer finding mask (a) taken from CBIS-DDSM dataset [32].

screening and treatment in the United States were \$1.08 billion and \$1.36 billion, respectively [21]. The National Breast and Cervical Cancer Early Detection Program (largest cancer screening program for low-income women in the United States) estimated that the total cost of all their services to be \$255.53 million during the 1-year period (2006-2007), as 1.02 million women were served [22]. In developing countries, such as Kenya, a full diagnostic test for BC can cost an average of \$401 in public facilities and \$1,205 in private facilities [23].

Diagnosis, management, and treatment. Women diagnosed with BC may experience cancer-related symptoms, such as: lump in the breast or under the arm (armpit), swelling or thickening part of the breast, irritation or dimpling of breast skin, redness or flaky skin in the nipple area or the breast, nipple discharge other than breast milk, including blood, changes in the size or the shape of the breast, and pain in the breast [24, 25]. The symptoms typically vary from patient to patient and some cases could even remain asymptomatic [26].

Early detection of BC is important to reduce the mortality rate, and it needs to be done when the patients are still asymptomatic [27, 28, 29]. Screening for BC is necessary so that the patient starts their treatment earlier, before cancer spreads to other tissues. Five-year survival rate of BC patients in North America is above 80% because of a more successful early detection of this disease [30]. In countries with limited resources, the situation is different and BC is commonly diagnosed at late stages [31].

2.2. Mammography

Image acquisition. Mammographic imaging produces an X-ray image of the breast. The image can be then used to check for BC in men and women with or without symptoms. Screening mammography routinely assesses the condition of a breast. If the results in the screening mammogram are suspicious or indicate signs of BC, a diagnostic mammogram is acquired to evaluate and diagnose unusual breast changes detected on a screening mammogram. The market around BC was estimated at $2.9 billion in 2016 [33].

The process of image acquisition can be described as follows. Acquisition of mammograms (X-rays images of breasts) is done by a trained technologist (radiographer), who positions the breast on a special platform in the imaging unit, and then gradually compresses the breast with a clear plastic paddle. Breast compression is necessary to even out the breast thickness, such that all the tissue can be visualized and small abnormalities are less likely to be hidden by overlying tissue. Spreading out the tissue also allows the use of a lower X-ray dose since a thinner tissue is being imaged. The patient must hold still for a few seconds while the X-ray picture is taken to reduce the possibility of a blurred image while the technologist walks behind a wall or into the next room to activate the X-ray machine. The routine views are a top-to-bottom view and an angled side view. The technologist will change the breast position between images. This process will be repeated for the other breast as well. The whole imaging process can take to about 30 minutes [5].

Quality of mammographic images. Proper exposure, contrast, resolution, breast compression and positioning are important for high quality mammographic images [34]. High quality images allow the radiologists to thoroughly examine mammograms and successfully detect present abnormalities [34]. After the introduction of digital mammography systems, the quality of mammograms has improved significantly [35]. With the new advancements in hardware and software that came with the digital systems, the factors affecting image quality like exposure, sharpness, noise, and contrast are being taken care of, but positioning is still visually assessed by technologists [35].

Breast positioning is one of the most important factors affecting a mammogram, as good positioning maximizes the amount of breast tissue detected in the imaging process [35, 36]. Breast pathology can be missed, since incorrect positioning may lead to image artifacts [37]. Reducing the overload on doctors by using artificial intelligence can lead to a better and faster diagnostics.

2.3. Quality Assessment in Mammography

PGMI system. There are various methods and systems that exist to evaluate and classify the quality of mammographic images. PGMI method is a four-grade ("Perfect", "Good", "Moderate", "Inadequate") system that was customized and adopted by many European countries for use in their BC screening programs [38]. "Perfect" means that: all breast tissues are imaged, correct image identification is clearly shown, correct exposure according to workplace requirements, good

Criteria for image assessment PGMI based	P = Perfect images	G = Good images	M = Moderate images (Acceptable for diagnostic purposes)	I = Inadequate images
1- All breast tissue imaged (fat tissue visualized posterior to glandular tissue) 2- Correct image identification clearly shown - Date of examination - Client identification—name and (number and/or date of birth) - Side markers - Positional markers - Radiographer identification 3- Correct exposure according to workplace requirements 4- Good compression 5- Absence of movement 6- Correct processing 7- Absence of artifacts 8- No skin folds 9- Symmetrical images	MLO images meet criteria for image assessment 1−9	- 1. all breast tissue imaged - Pectoral muscle well demonstrated - Nipple in profile - Infra-mammary fold (IMF) well demonstrated - 2 - 6. MLO images meet criteria for image assessment 2−6 inclusive for categorization as G - 7 - 9. MLO images displaying minor degrees of variation in criteria for - Imaging assessment 7, 8, and 9 will be accepted for categorization as G	- Most breast tissue imaged. - Pectoral muscle not to nipple level but posterior breast tissue adequately shown - Nipple not in profile but clearly distinguishable from retro-areolar tissue - IMF not clearly demonstrated but breast tissue adequately shown - Correct(ed) image identification - Correct exposure - Adequate compression - Absence of movement - Correct processing - Artefacts which do not obscure the image - Skin folds which do not obscure the breast tissue - Asymmetrical images	- A significant part of the breast not imaged - Incomplete or incorrect identification - Incorrect exposure - Inadequate compression hindering diagnosis - Blurred image - Incorrect processing - Overlying artifacts - Skin folds obscuring the image

Figure 4. PGMI quality assessment system.

compression, no movement, the images are processed correctly, no artifacts, no skin folds and the image is symmetrical. "Good" means: all breast tissues are imaged, correct image identification and it is clearly shown, correct exposure according to workplace requirements, good compression, absence of movement, correct processing, minor artifacts and skin folds with not fully symmetrical image. "Moderate" stands for: most breast tissue imaged, correct image identification, correct exposure, adequate compression, absence of movement, correct processing, artefacts which do not obscure the image, skin folds which do not obscure the breast tissue, asymmetrical images. "Inadequate" implies: a significant part of the breast is not imaged, incomplete or incorrect identification of the image, the exposure is incorrect, inadequate compression hindering diagnosis, blurred image, the images are processed incorrectly, overlying artifacts in the image, skin folds obscuring the image.

Automatic quality assessment: prior work. Most of the systems that are used to evaluate and grade images are reader-dependant. For this reason, the quality of a mammogram may be different if graded by two different readers [39]. To solve this problem an automated method using deep learning (DL) may be considered. Such method will evaluate and grade mammographic images without human interaction, and potentially provide an assessment quality similar to the one of a trained expert. A supervised DL-based framework can be used to classify breast tissue using Convolutional Neural Network to learn discriminate features automatically [40].

A proof of concept for an end-to-end DL framework to assesses image quality on the basis of single mammography images was developed using a database with known ground truth for training a regression CNN [41]. A DL algorithm was developed to assess mammographic breast density, showing results similar to the ones of experienced radiographer [42]. To automatically localize the nipple in mammograms a simple method was tested on 24 images, and the nipple was correctly located in 96% of the cases [43]. A fully automated method was described for pectoral muscle segmentation that is based on edge detection, intensity thresholding, surface smoothing, and straight line fitting [44].

In most of the previous studies DL was used to detect particular findings, but have never been used to segment different parts of the breast including skin-folds and do the quality assessment on the predicted masks similarly to the presented project.

2.4. Deep Learning

Deep learning for medical imaging. DL is a sub-field of machine learning [45], focusing on end-to-end representation learning from raw data [46]. DL, particularly convolutional neural networks, can be used to solve wide range of image classification tasks (assigning labels to images), and, furthermore, it is also highly promising in automatic segmentation of images (assigning a label to every image pixel) [47, 48, 49]. In many problems, state-of-the-art accuracy can be achieved, often exceeding human performance. DL models use labeled data and train using neural network architectures [50].

The most popular types of neural networks for computer vision applications are convolutional neural networks (CNN), with outstanding performance in machine learning problems [51]. CNNs have been proven to be very effective in areas such as image classification [52] and segmentation by learning features from the data. By using 2D convolutional layers, these architecture is well suited to processing 2D data like images [50]. CNNs mainly contain convolutional, activation, downsampling, and upsampling to perform operations over features, sometimes concluded with a number of fully connected layers. Fully convolution networks are basically CNNs with 1×1 convolutions instead of the dense layer.

During the recent years, DL has been getting a lot of attention because it is able to achieve breakthrough results in a variety of tasks: computer vision, speech recognition, and natural language processing [53, 50, 54]. The amount of available data plays a big role in DL [55]. DL also requires large computing power like high performance graphics processing units (GPU). GPU uses a parallel architecture that is efficient for DL. Speech recognition was between the first major application that was possible due to the advent of fast GPUs [50].

To date, machine learning has been applied to a wide range of applications [56]. In the medical imaging domain, DL showed superiority in many problems over other machine learning methods [57]. As such, DL has been very successful in computer vision applications [58, 59]. There is a lot of challenges related to using DL in medical settings, but the results that it gives are too valuable to discard [60], and many researchers focus on empirical evaluation of this methodology, including the present thesis.

Artificial neural networks. The human brain has hundreds of billions of cells called neurons [61]. The neurons are made of a cell body that is responsible for processing and carrying information towards and away from the brain [62]. Artificial neural networks (ANN) are function approximation machines guided by mathematical and engineering disciplines that are designed to obtain statistical generalization (and not to model the brain) [46]. The naming here is inspired from neuroscience [46]. A typical ANN has hundreds or thousands of artificial neurons, which consist of the weights and activation functions. Neuron's weights are multiplied by the input, and then transformed using an activation function. In an ANN setting, such neurons are interconnected, which allows them to be powerful models, able to automatically extract patterns from large datasets [53].

Optimization of neural networks. Most DL algorithms rely on some sort of optimization methods. These methods generally optimize the neural networks by minimizing a specific loss function [46, 63]. Loss function, in the context of machine learning, is used to show how good and bad the model is performing by measuring the errors made by a neural network [46, 63]. One particular method, used to optimize the loss of the neural network is a stochastic gradient descent (SGD) [64].

The most popular type of neural network optimization is done in a supervised setting. which is a subcategory of machine learning that uses labeled datasets to train the model [65]. Training data is used to compute the loss and its gradients with respect to the model's weights. These gradients are then used by SGD to update the model for a series of training iterations.

Overfitting and cross-validation. The goal of the training process is to maximize the prediction accuracy on the new data (test data). Overfitting happens when the model learns to perform well on the training data rather than generalize to new data [66]. One way to check that the model is overfitting is by comparing the losses of the training and validation datasets. The model is overfitting if the loss function on the training data will continue to show lower values in each epochs, but validation loss will eventually increase (Figure 5). In general, even simple linear models tend to overfit if the training set size is smaller than number of data dimensions[67].

Cross validation is a statistical method used to test the performance of machine learning models [68]. To conduct cross validation, data need to be split into two parts, where one part is used to train the model and the other is to validate the learning process. K-fold cross validation approach is considered a good approach to further improve the prediction accuracy [69]. It means splitting the data into k folds, train the model using $k - 1$ of them, and validate on the remaining part. The mean validation performance is then calculated after such runs (Figure 12). Cross validation is frequently used in machine learning to make a comparison and select models as it has lower bias than other methods [70].

Regularization Among the main problems related to DL, one is how to make an algorithm that performs well on the new data, and not just on the training data [46]. Different regularization techniques are used in training DL models to prevent overfitting. A model can be regularized by using weight decay, dropout, data augmentation, etc. Weight decay is a standard technique for regularizing models [71, 67, 72, 73]. One way to decrease complexity would be to add the weights to the loss function. To implement weight decay we add the squares of all the parameters to the loss function. This new term may get too large in relation to the main loss function, and to prevent that, we multiply the sum of squares with some value (weight decay) to make it smaller [71, 67]. A fully connected neural network, essentially, is a number of weights that are connected together that takes an input and gives prediction as an output. The prediction of the network is used to back-propagate the error to different

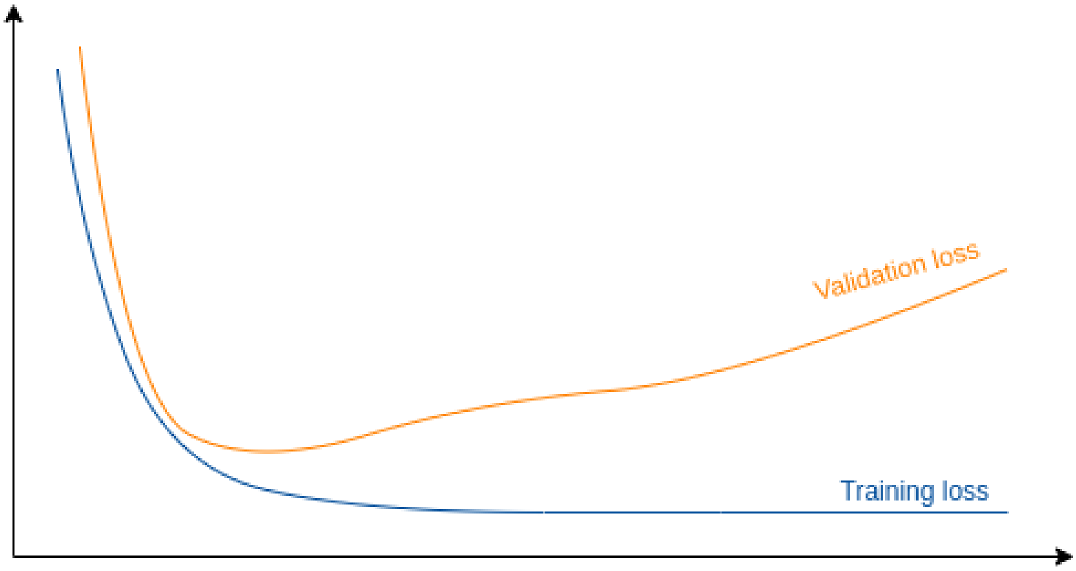


Figure 5. Example showing how overfitting looks like from training and validation losses.

layers and weights will be updated accordingly. If the weights in some part of the network are too big, they will have a greater impact on the predictions. This can lead the model to overfit. To address this problem, dropout is used to distribute the learnt representations more evenly across multiple weights such that the model does not overfit [72, 71]. Data augmentation is another technique used to regularize the model training and help in preventing the overfitting [73]. It is based on artificial generation of the images based on existing samples from the training dataset via small random perturbations applied to them. Creating variations of the images can improve the ability of the model to generalize the prediction to new images [74]. Image augmentation is used to tackle overfitting problem and to increase of the existing dataset [73, 75]. DL network models can give better results when trained on more data [46, 67]. Images are modified through different ways of processing, such as random rotation, shifts, shear and flips, etc.

Transfer learning approach. Transfer learning is using the knowledge of an already trained (pre-trained) model on a new but related problem. In machine learning a model developed for a task can be reused in a second model as a starting point. Pre-trained models are a popular approach in DL as they provide a huge increase in performance in similar problems, particulaly, in case of a small dataset [76].

Deep learning for image segmentation. Image segmentation is a common problem in computer vision and digital image processing [77]. The purpose of image segmentation is to process and localize the pixels of objects and boundaries, converting digital images into different classes that are easier to analyze by computers [78, 77]. As a result, a set of segments that is covering the image or a set of contours is extracted from an image. After segmenting the image, further analyses, e.g. object detection,

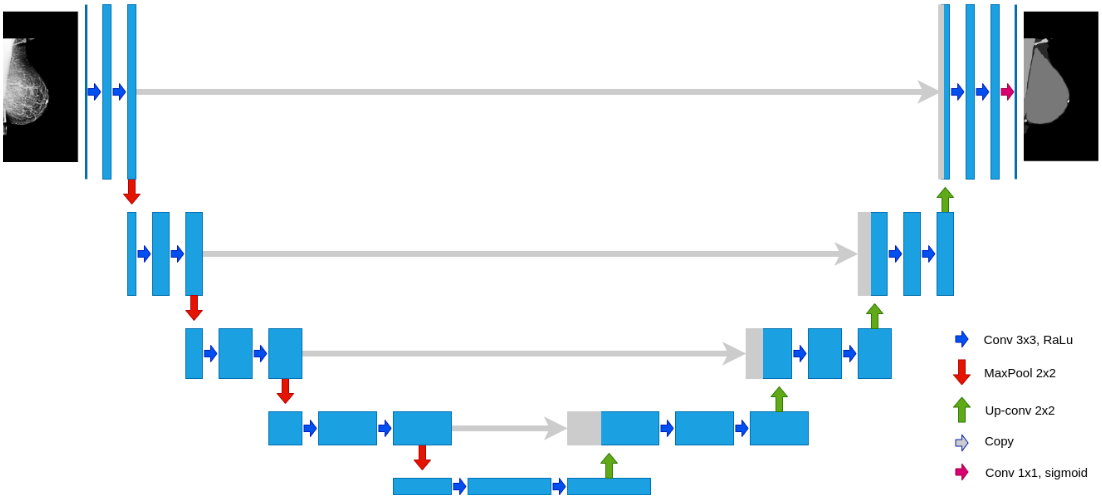


Figure 6. Example of U-Net architecture.

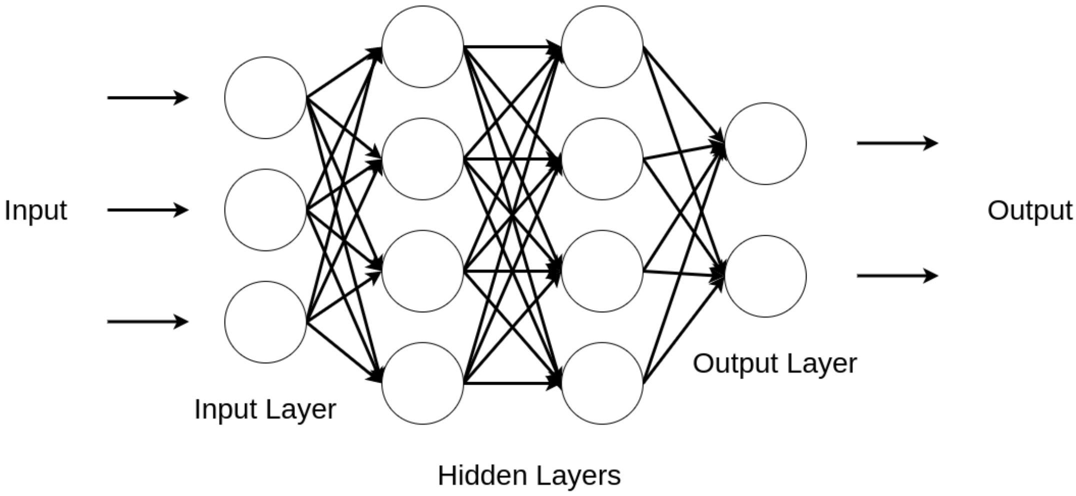


Figure 7. Example of a simple fully connected neural network with an input, an output, and 2 hidden layers.

can be performed [78]. In medical imaging, image segmentation is used to extract clinically relevant information from the images [79].

Numerous studies on segmentation of medical images have been introduced in the past [80, 81, 49]. In the context of BC, a novel method was developed for breast masses segmentation from mammograms using DL, to show the accuracy and efficiency that these techniques can provide [82]. An integrated computer-aided diagnosis system was used to screen digital X-ray mammograms for detection, segmentation, and classification of breast masses using DL methods [83]. A fully convolutional neural network (FCNN) with attention-guided deep supervision was developed for automatic segmentation, improving the overall segmentation accuracy [84]. The present work is also applying image segmentation to mammography and it uses DL.

3. OBJECTIVES OF THE STUDY

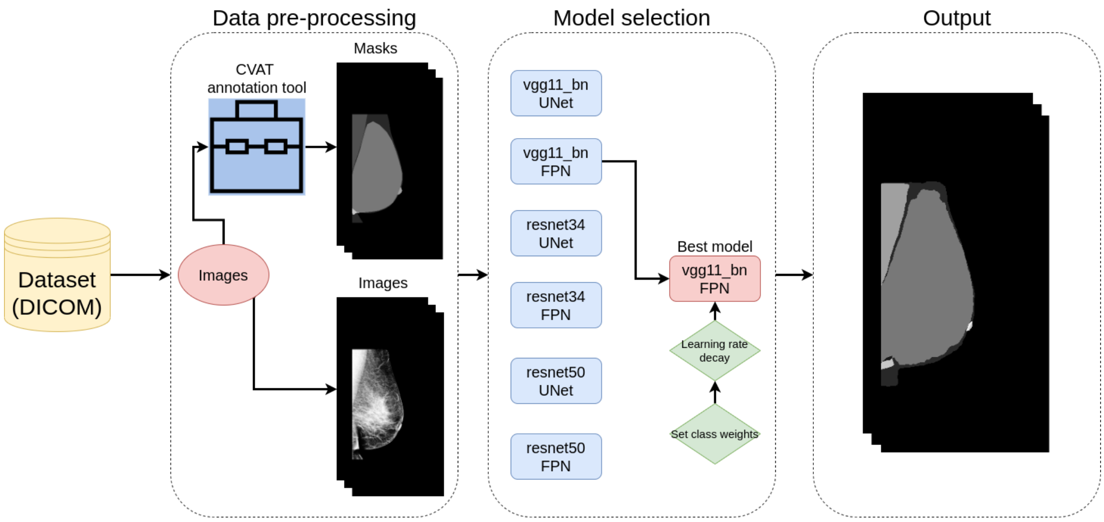


Figure 8. The workflow of the segmentation process designed in the present thesis.

Graphical illustration of the workflow executed in this thesis is provided in Figure 8. The main objective of this study was to develop a proof-of-concept of automated mammography quality assessment system via image segmentation. The study also had two sub-objectives:

- Create (collect and annotate) a dataset for mammography tissue and artifact segmentation that is representative in the scope of the considered quality assessment criteria (see Section 2.3);
- Benchmark different existing Deep Neural Network architectures and select the most appropriate one for the segmentation of defects in mammograms.

4. METHODS AND MATERIALS

4.1. Data Collection and Annotation

Data source and pre-selection. The author used the data from the Oulu University Hospital (approved by the institutional review board). The dataset consisted of 503 mammography images. The original dataset consisted of 251,422 mammography studies, where each study included Cranio-caudal (CC) view and Mediolateral Oblique (MLO) view images. In this project, we focused on MLO images only. The dataset was in Digital Imaging and Communications in Medicine (DICOM) format, so multiple scripts were developed to process and extract 5,000 random images and convert them to Portable Network Graphics (PNG) format along with the corresponding patient identification numbers. That information were saved in separate comma-separated values (CSV) file. 1,000 images with bad positioning and artifacts (i.e. skin folds) were chosen manually by the author from the original 5,000 images dataset.

The original images were large and had variability in size, which could affect the deep learning models. Furthermore, large images need more computation power and time to run the pipeline. A script was developed to resize all images and masks (described below) to 512×256 pixels. The script was made such that the resulting images had the aspect ratio as the original one, achieved by padding the sides with zeros if needed (Figure 11). The resized images and masks were faster to train on than the data of original size. The images and masks processed this way were used to train and evaluate the segmentation models.

Data annotation A selected subset of 503 images was annotated manually using the CVAT annotation tool [85] (Figure 9). Five different pixel classes were annotated in each image (Figure 10). Importantly, not all classes were presented in all the images. The three major classes (whole breast, breast, and muscle) existed in the vast majority of the images. Other classes, such as skin-folds and nipple were present less, but in most of the images as well. Skin-folds is different from other classes in the case that there can be more then one skin fold in an image. The annotation process of one image took between 2-6 minutes depending on the mammogram. The annotations were, eventually, stored in JavaScript Object Notation (JSON) format for further processing.

4.2. Experimental Setup

Data processing. All the experiments were done using Pytorch [86] and Pytorch-lightning. The selected 503 images were used in our experiments, with the data were split to training images (n=321), validation (n=81), and testing (n=101). We employed light data augmentations. For training we used padding to 520×270 pixels, random horizontal fliping, and cropping to 512×256 pixels. For validation and testing we did not use any augmentations.

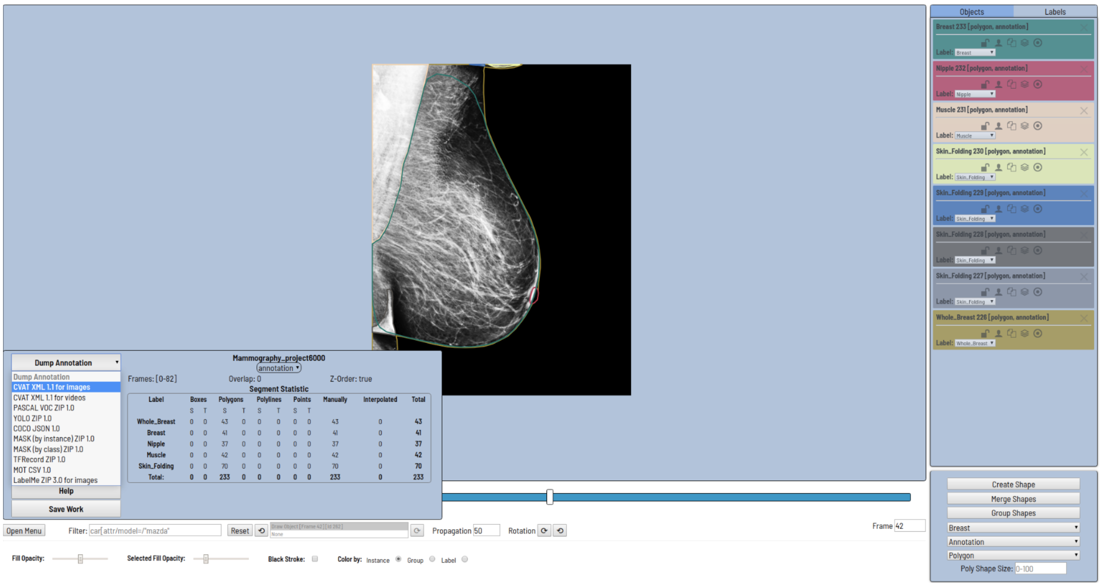


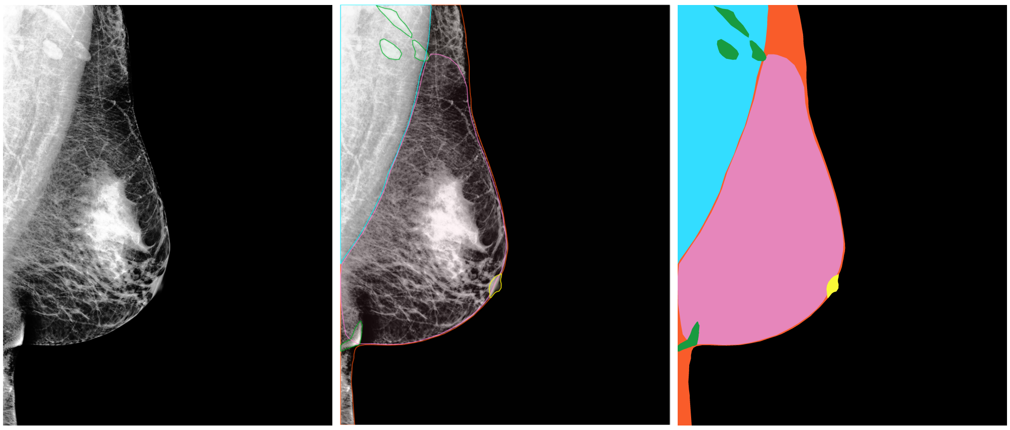
Figure 9. CVAT annotation tool: The picture shows the user interface of the tool. Here, the wanted class is chosen by clicking on Breast in the right down corner. Then, the annotation is done by clicking on Create Shape and draw a polygon around the class.

Model selection. The performance of VGG11-BN [87] with U-NET [88], VGG11-BN with FPN [89], RESNET34 [90] with U-NET, RESNET34 with FPN, RESNET50 [90] with U-NET, and RESNET50 with FPN decoders were investigated.

U-net is a network and training strategy that was originally developed for medical image segmentation [88]. The architecture consists of two paths, encoder (contraction) and decoder (expansion). The encoder is made of convolutional and max pooling layers used to capture the context in the image. The decoder is symmetric expanding path, which use transposed convolutions to enable precise localization (Figure 6). It relies on the use of data augmentation to utilize the annotated samples more efficiently. The architecture consists of a contracting path to capture context and a symmetric expanding path that enables precise localization. Feature Pyramid Networks (FPN) is a feature extractor network designed for image segmentation. The network is designed based on representation pyramid concept that gives better accuracy and speed. It takes a single-scale image as input and outputs feature maps at multiple levels [89].

Group K-fold cross-validation was implemented on the training data, see Figure 12. This procedure ensured that data from the same patients and studies were not present in training and validation splits simultaneously in any of the folds. Out-of-fold (OOF) script was made to get the predictions from each set of folds and then average them to assess the performance on validation.

The models were optimized by minimizing the multi-class cross-entropy loss function. Dice coefficient was used to evaluate the performance of the models and choose the best model at validation. Dice[91] coefficient have been used extensively for the assessment of image segmentation algorithms . Dice coefficient measure of how many positives the algorithm can find and penalizes for the false positives (Figure 13).



(a) Original image (b) Overalyed countours (c) Masks

Figure 10. Example of the different classes plotted in different colors, that were made and used in the segmentation process.

(a) (b)

Figure 11. Example of original and resized image – (a) and (b), respectively. In the subfigure (b) the image was resized and repositioned so that the aspect ratio is similar to the original one, by padding the sides when needed.

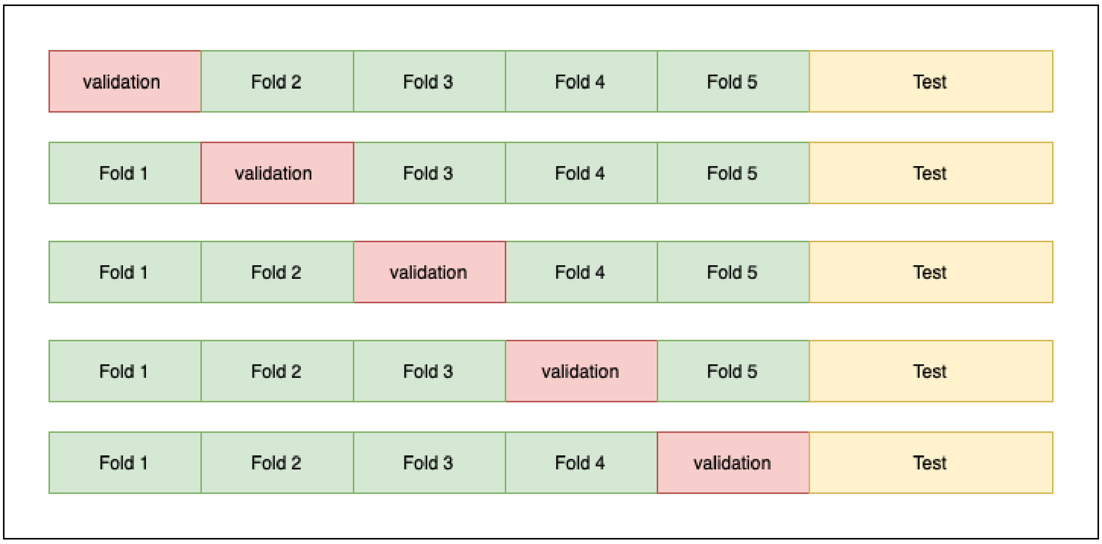


Figure 12. Different folds used in 5-folds cross-validation setting.

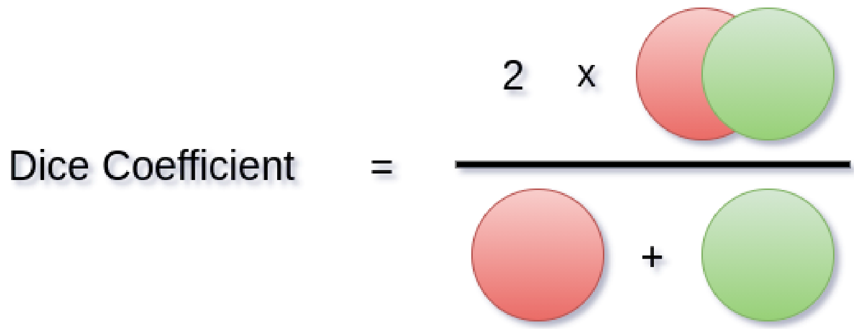


Figure 13. Illustration of Dice Coefficient.

4.3. Quality Assessment

The author received a proper introduction into breast anatomy, breast cancer, mammography, and mammography quality assessment. At the moment of the study execution, the topics of breast segmentation and quality assessment of mammograms were covered in the literature very sparsely and in a non-systematic way. The introduced approach to automate quality assessment of mammography images is novel. As an primary requirement, the annotated data was needed to train deep learning models to segment different parts of the breast automatically.

The mammography quality assessment criteria for this project were taken from "Early detection of breast cancer for health professionals" course [6]. Different criteria were chosen to make a quality assessment script. A script was developed to check the image quality according to the quality criteria (Figure 14). The script used the predicted masks (from the testing set) to check the selected quality criteria like presence of nipple and skin-folds, muscle length, and muscle angle, and report a list of the these four criteria (Figure 16). The predicted masks from JSON files were used as independent classes that may overlap.

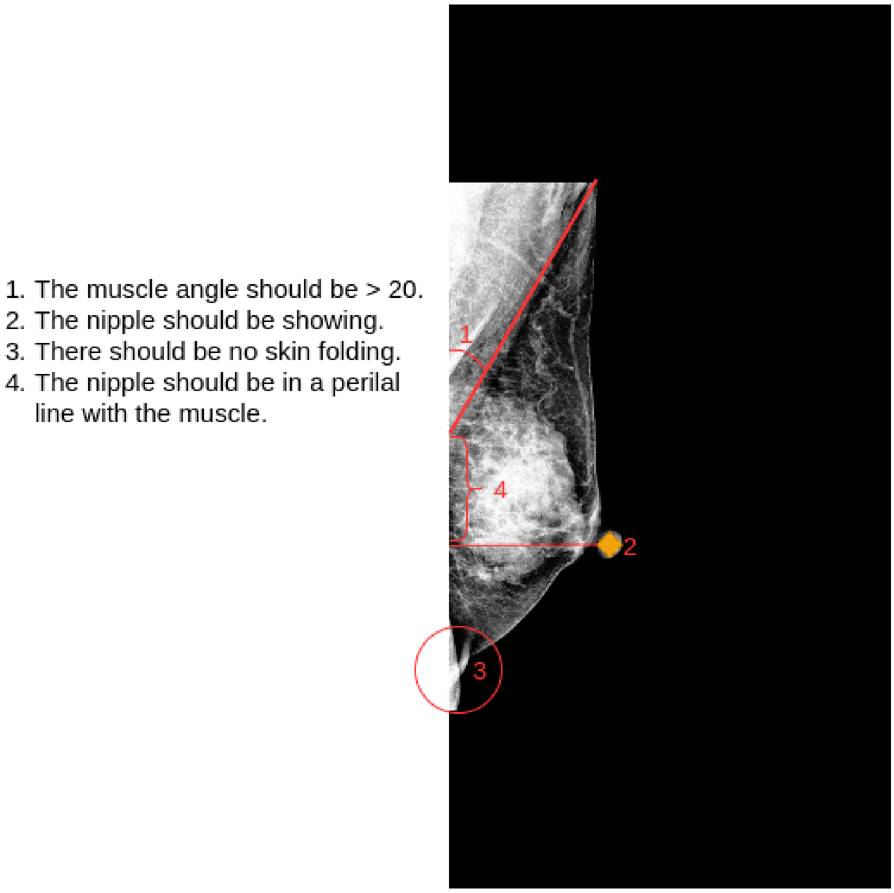


Figure 14. Quality assessment criteria implemented in this project.

5. RESULTS

5.1. Optimal Backbone-Decoder Configuration

Deep learning-based segmentation models achieved high Dice score results in segmenting different classes. From Table 1, FPN with Vgg11-bn was chosen to be the best model, as the Dice score of the model was better then with the other models.

Decoder	Backbone	Whole Breast	Breast	Muscle	Skin Folding	Nipple
U-Net	Vgg11_bn	$0{,}985 \pm 0{,}011$	$0{,}969 \pm 0{,}006$	$0{,}975 \pm 0{,}009$	$0{,}542 \pm 0{,}051$	$0{,}456 \pm 0{,}050$
U-Net	Resnet34	$0{,}985 \pm 0{,}011$	$0{,}969 \pm 0{,}005$	$0{,}975 \pm 0{,}010$	$0{,}536 \pm 0{,}051$	$0{,}476 \pm 0{,}054$
U-Net	Resnet50	$0{,}985 \pm 0{,}011$	$0{,}972 \pm 0{,}005$	$0{,}976 \pm 0{,}009$	$0{,}537 \pm 0{,}052$	$0{,}478 \pm 0{,}054$
FPN	**Vgg11_bn**	$\mathbf{0{,}985 \pm 0{,}011}$	$\mathbf{0{,}974 \pm 0{,}004}$	$\mathbf{0{,}982 \pm 0{,}005}$	$\mathbf{0{,}641 \pm 0{,}052}$	$\mathbf{0{,}586 \pm 0{,}051}$
FPN	Resnet34	$0{,}985 \pm 0{,}011$	$0{,}974 \pm 0{,}005$	$0{,}982 \pm 0{,}004$	$0{,}623 \pm 0{,}051$	$0{,}573 \pm 0{,}048$
FPN	Resnet50	$0{,}985 \pm 0{,}011$	$0{,}975 \pm 0{,}004$	$0{,}979 \pm 0{,}005$	$0{,}628 \pm 0{,}051$	$0{,}570 \pm 0{,}048$

Table 1. Dice score (2 * area of overlap / total number of pixels) results for different architectures. Numbers shown are mean and standard deviation of Dice scores over the testing data.

5.2. Best Model Assessment

The best model was modified the retrained to get better results, see Table 2. Whole breast, muscle, and breast classes were showing Dice score larger than 90% due to the existence of these classes in most of the images and the similarities of the classes like place and shape in different mammograms. Needless to say that the relatively small size of the dataset effected on the results of this project. Overall, the pipeline was successfully able to segment all the classes, see Figure 15. The training took around 30 minutes for one model, and repeated 5 times (because of choosing 5 folds for cross-validation).

Quality assessment software was implemented to check the classes of the predicted masks from the testing set. The script was able to classify the masks according to the chosen quality criteria (Figure 14) and print the results as a list of the these four criteria (Figure 16).

Decoder Backbone	FPN VGG11_BN
Whole Breast	$0{,}972 \pm 0{,}011$
Breast	$0{,}943 \pm 0{,}008$
Muscle	$0{,}970 \pm 0{,}008$
Skin Folding	$0{,}641 \pm 0{,}052$
Nipple	$0{,}634 \pm 0{,}048$

Table 2. Dice score results after retraining the best architecture.

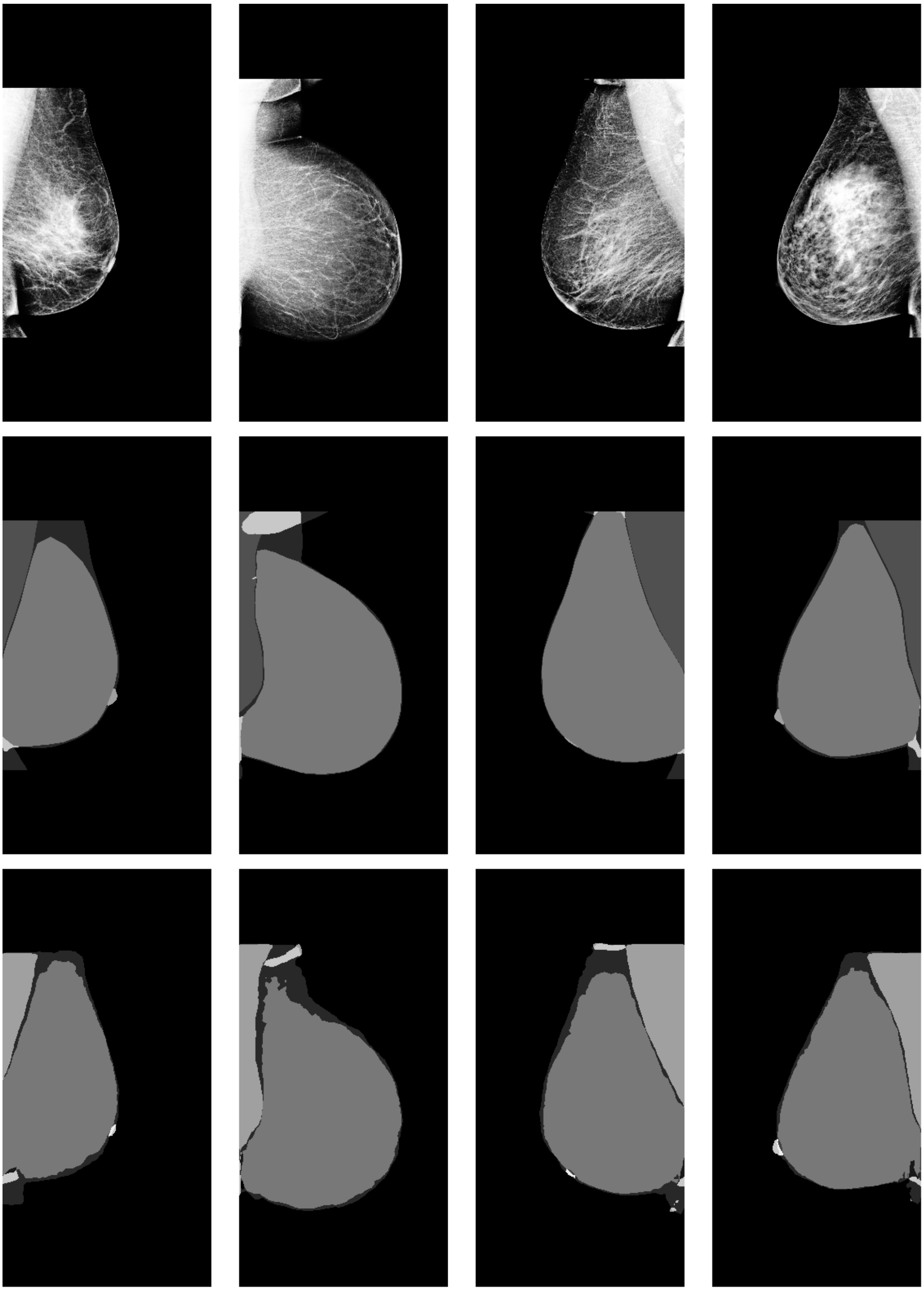


Figure 15. Examples of the segmentation results. The first row shows the original images. The second row shows the manual segmentations. The third row shows the prediction of our pipeline.

Image Number is 87
- The nipple is in profile
- The pectoral muscle is sufficiently long: in the same level to the nipple
- There is skinfolds in the image
- The muscle angel is 11.06

Figure 16. The quality assessment results automatically generated from the segmentation masks using the developed software tool.

6. DISCUSSION

Mammography has many challenges, from positioning [35, 36] to detection of breast abnormalities and presence of overlapping tissues [92]. In this project our focus was on assessment of mammography positioning. The author performed a feasibility study and implemented an automatic approach for quality assessments in mammography screening using deep learning segmentation model. Deep learning based segmentation methods was used to do this task.

There was no publicly available dataset to be used in this project. In order to solve this problem, a mammography dataset (images, masks) was manually created by the author based on the extracted mammography studies from Oulu University Hospital. Preprocessing scripts were developed to extract the images from DICOM files with corresponding patient information.

The results indicate that different breast tissue and findings can be segmented, and the quality assessment tool is then able to use the segmented regions to evaluate breast positioning in mammograms automatically. The study provides a new automatic application to solve the positioning problem in mammographic imaging.

When making the dataset, the author chose the presumably worst images to segment (according to PGMI criteria), so to diversify the training data. That led to choose mammograms where there was no nipple because nipple presence is one of the criteria for a good and diagnostic mammogram. As a result, nipple class was the hardest for the model to segment due to the lack of nipple in training data and, additionally, due to the small size of the nipple area. Segmenting skin folding was also hard due to different patterns of the folds. Moreover, unlike other considetred classes that have relatively similar positions in mammograms, skin foldings had larger variations in terms of location. Manual segmentation of skin folding was also ambiguous, since edges of skin foldings are not clearly defined.

Despite the above, the created dataset still has a clear value for segmentation and localization purposes, and the developed segmentation pipeline is likely to be applicable for other segmentation application that include multi-class segmentation.

To improve the segmentation accuracy, other preprocessing methods can be tested like trying different sizes of the images and smoothing the images. Also, testing of different methods for preprocessing, data augmentation, and patch-based sampling of medical images could be investigated [93]. Several medical studies used segmentation with multiple combinations of preprocessing methods such as noise removal, histogram equalization, and edge enhancement to enhance the quality of input images [94]. Different augmentation strategies were compared, showing how much augmentation can affect model performance in medical images [95].

More development on the quality assessment can be done, so to increase the completeness of the quality assessment criteria that the software evaluates, such as, for example, the blurriness of a breast. A deep learning based software can be applied to automatically detect and segment blurry areas in whole slide images [96]. An algorithm can be applied to restore a clear high-resolution from blurry low-resolution images [97].

7. CONCLUSIONS

To conclude, in this thesis, the author developed a comprehensive method for mammography image segmentation. From the methodological point of view, the author found that the use of VGG11 backbone and FPN decoder in an encoder-decoder model is well suited for this task. The results show that segmentation of structures with not well-defined shape and high heterogeneity in appearance, such as skin foldings, is also feasible with the developed model. Potential applications of the method are in automatic quality assurance in clinical practice.

8. ACKNOWLEDGEMENT

I would like to acknowledge Oulu University Hospital for providing the mammography studies. Research Unit of Medical Imaging, Physics and Technology, University of Oulu, is acknowledged for providing the infrastructure for this project. Dr. Aleksei Tiulpin and Mr. Egor Panfilov are acknowledged for supervising this study. Mr. Antti Isosalo is acknowledged for fruitful discussions related to mammography. Mr. Huy Hoang Nguyen is acknowledged for providing guidance related to deep learning.

BIBLIOGRAPHY

[1] J. Ferlay et al. ”Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods”. In: *International Journal of Cancer* 144.8 (2019), pp. 1941–1953. DOI: 10.1002/ijc.31937.

[2] ECIS. *European Cancer Information System*. URL: https://ecis.jrc.ec.europa.eu/?Cancer=46&Gender=2. (2020).

[3] Ying Wang et al. ”Mammographic mass segmentation: Embedding multiple features in vector-valued level set in ambiguous regions”. In: *Pattern Recognition* 44.9 (2011). Computer Analysis of Images and Patterns, pp. 1903–1915. ISSN: 0031-3203. DOI: https://doi.org/10.1016/j.patcog.2010.08.002.

[4] San-Kan Lee et al. ”A computer-aided design mammography screening system for detection and classification of microcalcifications”. In: *International Journal of Medical Informatics* 60.1 (2000), pp. 29–57. ISSN: 1386-5056. DOI: https://doi.org/10.1016/S1386-5056(00)00067-8.

[5] Radiology Info. *Mammography*. Ed. by Radiologyinfo.org. radiologyinfo.org. Apr. 2019.

[6] *Early detection of breast cancer for health professionals course*. URL: http://earlydetectionofbreastcanser.weebly.com/.

[7] Pan American Health Organization. ”Mammography Services Quality Assurance: Baseline Standards for Latin America and the Caribbean”. In: *PLOS ONE* (Dec. 2016). URL: https://iris.paho.org/handle/10665.2/31402.

[8] Eija Metsälä et al. ”European radiographers’ challenges from mammography education and clinical practice – an integrative review”. In: *Insights into Imaging* 8 (2017), pp. 329–343. DOI: 10.1007/s13244-016-0542-1.

[9] B. Strøm et al. ”Challenges in mammography education and training today: The perspectives of radiography teachers/mentors and students in five European countries”. In: *Radiography* 24.1 (2018), pp. 41–46. ISSN: 1078-8174. DOI: https://doi.org/10.1016/j.radi.2017.08.008.

[10] Wang Kei Ma et al. ”Blurred digital mammography images: an analysis of technical recall and observer detection performance”. In: *The British Journal of Radiology* 90.1071 (2017), p. 20160271. DOI: 10.1259/bjr.20160271.

[11] Vairaprakash Gurusamy, Subbu Kannan, and G.Nalini. ”Review on Image Segmentation Techniques”. In: Oct. 2014.

[12] Feng Y et al. ”Breast cancer development and progression: Risk factors, cancer stem cells, signaling pathways, genomics, and molecular pathogenesis”. In: *Genes Dis* (May 2018). URL: doi:10.1016/j.gendis.2018.05.001.

[13] Yalaza M, İnan A, and Bozer M. ”Male Breast Cancer”. In: *J Breast Health* (Jan. 2016). URL: `doi:10.5152/tjbh.2015.2711`.

[14] Emin Yıldırım and Uğur Berberoğlu. ”Male breast cancer: a 22-year experience”. In: *European Journal of Surgical Oncology (EJSO)* 24.6 (1998), pp. 548–552. ISSN: 0748-7983. DOI: `https://doi.org/10.1016/S0748-7983(98)93608-3`.

[15] SONALI PANDYA and RICHARD G. MOORE. ”Breast Development and Anatomy”. In: *Clinical Obstetrics and Gynecology* 54 (2011). URL: `https://journals.lww.com/clinicalobgyn/Fulltext/2011/03000/`.

[16] Akram M et al. ”Awareness and current knowledge of breast cancer”. In: *Biol Res* (Oct. 2017). DOI: `doi:10.1186/s40659-017-0140-9`.

[17] Chuba PJ et al. ”Bilateral risk for subsequent breast cancer after lobular carcinoma-in-situ: analysis of surveillance, epidemiology, and end results data. J Clin Oncol”. In: *PMID* (Aug. 2005). DOI: `https://doi:10.1200/JCO.2005.04.038.PMID:16110014`.

[18] Karpanen T and Alitalo K. *Lymphatic vessels as targets of tumor therapy?* Apr. 2001. DOI: `https://doi:10.1084/jem.194.6.f37`.

[19] Patrick J. Lynch. *Breast normal anatomy cross-section*. 2007. URL: `https://commons.wikimedia.org/wiki/File:Breast_anatomy_normal_scheme.png`.

[20] Rebecca L. Siegel, Kimberly D. Miller, and Ahmedin Jemal. ”Cancer statistics, 2017”. In: *CA: A Cancer Journal for Clinicians* 67.1 (2017), pp. 7–30. DOI: `10.3322/caac.21387`.

[21] Cary P. Gross et al. ”The Cost of Breast Cancer Screening in the Medicare Population”. In: *JAMA Internal Medicine* 173.3 (Feb. 2013), pp. 220–226. ISSN: 2168-6106. DOI: `10.1001/jamainternmed.2013.1397`.

[22] Donatus U. Ekwueme et al. ”Cost of services provided by the National Breast and Cervical Cancer Early Detection Program”. In: *Cancer* 120.S16 (2014), pp. 2604–2611. DOI: `https://doi.org/10.1002/cncr.28816`.

[23] Subramanian S et al. ”Cost and affordability of non-communicable disease screening, diagnosis and treatment in Kenya: Patient payments in the private and public sectors”. In: (2018). DOI: `https://doi.org/10.1371/journal.pone.0190113`.

[24] Minjoung Monica Koo et al. ”Typical and atypical presenting symptoms of breast cancer and their associations with diagnostic intervals: Evidence from a national audit of cancer diagnosis”. In: *Cancer Epidemiology* 48 (2017), pp. 140–146. ISSN: 1877-7821. DOI: `https://doi.org/10.1016/j.canep.2017.04.010`.

[25] Laura E. Davis et al. ”Patient-reported symptoms after breast cancer diagnosis and treatment: A retrospective cohort study”. In: *European Journal of Cancer* 101 (2018), pp. 1–11. ISSN: 0959-8049. DOI: `https://doi.org/10.1016/j.ejca.2018.06.006`.

[26] Majeed W et al. ”Breast cancer: major risk factors and recent developments in treatment”. In: *National Center for Biotechnology Information* (2014). DOI: http://dx.doi.org/10.7314/apjcp.2014.15.8.3353.

[27] Mugahed A Al-antari et al. ”An Automatic Computer-Aided Diagnosis System for Breast Cancer in Digital Mammograms via Deep Belief Network”. In: *Journal of Medical and Biological Engineering* 38 (2018). DOI: 10.1007/s40846-017-0321-6.

[28] Ryohei Takahashi and Yuya Kajikawa. ”Computer-aided diagnosis: A survey with bibliometric analysis”. In: *International Journal of Medical Informatics* 101 (2017), pp. 58–67. ISSN: 1386-5056. DOI: https://doi.org/10.1016/j.ijmedinf.2017.02.004.

[29] Yi-Sheng Sun et al. ”Risk Factors and Preventions of Breast Cancer”. In: (2017). DOI: 10.7150/ijbs.21635.

[30] Carol E. DeSantis et al. ”Breast cancer statistics, 2015: Convergence of incidence rates between black and white women”. In: *CA: A Cancer Journal for Clinicians* 66.1 (2016), pp. 31–42. DOI: 10.3322/caac.21320.

[31] Anderson Benjamin O. et al. ”Early Detection of Breast Cancer in Countries with Limited Resources”. In: *The Breast Journal* 9.s2 (2003), S51–S59. DOI: 10.1046/j.1524-4741.9.s2.4.x.

[32] Rebecca Sawyer Lee et al. ”A curated mammography data set for use in computer-aided detection and diagnosis research”. In: *Scientific Data* 4 (2017), pp. 1941–1953. DOI: 10.1038/sdata.2017.177.

[33] Grand View Research. *Breast Imaging Market Size, Share & Trend Analysis Report By Technology (Ionizing Technology, Non-Ionizing Technology), And Segment Forecasts, 2018 - 2024*. URL: https://www.grandviewresearch.com/industry-analysis/breast-imaging-equipment-market. (2018).

[34] G W Eklund, G Cardenosa, and W Parsons. ”Assessing adequacy of mammographic image quality.” In: *Radiology* 190.2 (1994), pp. 297–307. DOI: 10.1148/radiology.190.2.8284372.

[35] Manju Bala Popli et al. ”Breast Positioning during Mammography: Mistakes to be Avoided”. In: *Breast Cancer: Basic and Clinical Research* 8 (2014), BCBCR.S17617. DOI: 10.4137/BCBCR.S17617.

[36] Lawrence Bassett et al. ”Mammographic positioning: Evaluation from the view box”. In: *Radiology* 188 (Oct. 1993), pp. 803–6. DOI: 10.1148/radiology.188.3.8351351.

[37] Eren D. Yeh et al. ”Positioning in Breast MR Imaging to Optimize Image Quality”. In: *RadioGraphics* 34.1 (2014), E1–E17. DOI: 10.1148/rg.341125193.

[38] Ammar Haouimi, Stefano Pacifici, and et al. ”PGMI evaluation system”. In: *Radiopaedia.org* (). URL: https://radiopaedia.org/articles/pgmi-evaluation-system.

[39] M. Boyce et al. ”Comparing the use and interpretation of PGMI scoring to assess the technical quality of screening mammograms in the UK and Norway”. In: *Radiography* 21.4 (2015). Issues in breast imaging, pp. 342–347. ISSN: 1078-8174. DOI: https://doi.org/10.1016/j.radi.2015.05.006.

[40] A. Dubrovina et al. ”Computational mammography using deep neural networks”. In: *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization* 6.3 (2018), pp. 243–247. DOI: 10.1080/21681163.2015.1131197.

[41] T. Kretz et al. ”Mammography Image Quality Assurance Using Deep Learning”. In: *IEEE Transactions on Biomedical Engineering* (2020), pp. 1–1. DOI: 10.1109/TBME.2020.2983539.

[42] Constance D. Lehman et al. ”Mammographic Breast Density Assessment Using Deep Learning: Clinical Implementation”. In: *Radiology* 290.1 (2019). PMID: 30325282, pp. 52–58. DOI: 10.1148/radiol.2018180694.

[43] R. Chandrasekhar and Y. Attikiouzel. ”A simple method for automatically locating the nipple on mammograms”. In: *IEEE Transactions on Medical Imaging* 16.5 (1997), pp. 483–494.

[44] S. M. Kwok, R. Chandrasekhar, and Y. Attikiouzel. ”Automatic pectoral muscle segmentation on mammograms by straight line estimation and cliff detection”. In: *The Seventh Australian and New Zealand Intelligent Information Systems Conference, 2001*. 2001, pp. 67–72.

[45] F. Ertam and G. Aydın. ”Data classification with deep learning using Tensorflow”. In: *2017 International Conference on Computer Science and Engineering (UBMK)*. 2017, pp. 755–758. DOI: 10.1109/UBMK.2017.8093521.

[46] Ian Goodfellow et al. *Deep learning*. Vol. 1. 2. MIT press Cambridge, 2016.

[47] Shidan Wang et al. ”Pathology Image Analysis Using Segmentation Deep Learning Algorithms”. In: *The American Journal of Pathology* 189.9 (2019), pp. 1686–1698. ISSN: 0002-9440. DOI: https://doi.org/10.1016/j.ajpath.2019.05.007.

[48] Moeskops Pim et al. ”Deep Learning for Multi-task Medical Image Segmentation in Multiple Modalities”. In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*. Ed. by Ourselin Sebastien et al. Cham: Springer International Publishing, 2016, pp. 478–486.

[49] Matthew Lai. ”Deep Learning for Medical Image Segmentation”. In: *CoRR* abs/1505.02000 (2015). URL: http://arxiv.org/abs/1505.02000.

[50] Yann LeCun, Y. Bengio, and Geoffrey Hinton. ”Deep Learning”. In: *Nature* 521 (May 2015), pp. 436–44. DOI: 10.1038/nature14539.

[51] S. Albawi, T. A. Mohammed, and S. Al-Zawi. ”Understanding of a convolutional neural network”. In: *2017 International Conference on Engineering and Technology (ICET)*. 2017, pp. 1–6. DOI: 10.1109/ICEngTechnol.2017.8308186.

[52] Rajat Kumar Sinha, Ruchi Pandey, and Rohan Pattnaik. *Deep Learning For Computer Vision Tasks: A review*. 2018. arXiv: 1804.03928 [cs.CV].

[53] Jürgen Schmidhuber. "Deep learning in neural networks: An overview". In: *Neural networks* 61 (2015), pp. 85–117.

[54] Shixia Liu et al. "Towards better analysis of machine learning models: A visual analytics perspective". In: *Visual Informatics* 1.1 (2017), pp. 48–56. ISSN: 2468-502X. DOI: https://doi.org/10.1016/j.visinf.2017.01.006.

[55] Maryam Najafabadi et al. "Deep learning applications and challenges in big data analytics". In: *Journal of Big Data* 2 (Dec. 2015). DOI: 10.1186/s40537-014-0007-7.

[56] Arel Itamar, Rose Derek, and Karnowski Thomas. "Deep Machine Learning - A New Frontier in Artificial Intelligence Research [Research Frontier]". In: *IEEE Computational Intelligence Magazine* 5.4 (2010), pp. 13–18.

[57] Hyun-Jong Jang and Kyung-Ok Cho. "Applications of deep learning for the analysis of medical data". In: 42 (June 2019). DOI: 10.1007/s12272-019-01162-9.

[58] Athanasios Voulodimos et al. "Deep Learning for Computer Vision: A Brief Review". In: (2018). DOI: 10.1155/2018/7068349.

[59] Alex Krizhevsky, Sutskever Ilya, and Hinton Geoffrey E. "ImageNet Classification with Deep Convolutional Neural Networks". In: *Advances in Neural Information Processing Systems 25*. Ed. by F. Pereira et al. Curran Associates, Inc., 2012, pp. 1097–1105. URL: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf.

[60] Alexander Selvikvåg Lundervold and Arvid Lundervold. "An overview of deep learning in medical imaging focusing on MRI". In: *Zeitschrift für Medizinische Physik* 29.2 (2019). Special Issue: Deep Learning in Medical Physics, pp. 102–127. ISSN: 0939-3889. DOI: https://doi.org/10.1016/j.zemedi.2018.11.002.

[61] Suzana Herculano-Houzel. "The Human Brain in Numbers: A Linearly Scaled-up Primate Brain". In: *Frontiers in human neuroscience* 3 (Nov. 2009), p. 31. DOI: 10.3389/neuro.09.031.2009.

[62] Christof Koch and Idan Segev. "The role of single neurons in information processing". In: *Nature Neuroscience* (Nov. 2000). DOI: 10.1038/81444.

[63] Sathya R Chitturi et al. "Perspective: new insights from loss function landscapes of neural networks". In: *Machine Learning: Science and Technology* 1.2 (Apr. 2020), p. 023002. DOI: 10.1088/2632-2153/ab7aef. URL: https://doi.org/10.1088/2632-2153/ab7aef.

[64] Ilya Sutskever et al. "On the importance of initialization and momentum in deep learning". In: *Proceedings of the 30th International Conference on Machine Learning*. Ed. by Sanjoy Dasgupta and David McAllester. Vol. 28. Proceedings of Machine Learning Research 3. Atlanta, Georgia, USA: PMLR, June 2013, pp. 1139–1147. URL: `http://proceedings.mlr.press/v28/sutskever13.html`.

[65] Pádraig Cunningham, Matthieu Cord, and Sarah Jane Delany. "Supervised Learning". In: *Machine Learning Techniques for Multimedia: Case Studies on Organization and Retrieval*. Ed. by Matthieu Cord and Pádraig Cunningham. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 21–49. ISBN: 978-3-540-75171-7. DOI: `10.1007/978-3-540-75171-7_2`.

[66] Thomas G. Dietterich. "Overfitting and undercomputing in machine learning". In: *ACM Comput. Surv.* 27 (1995), pp. 326–327.

[67] Aston Zhang et al. *Dive into Deep Learning*. `https://d2l.ai`. 2020.

[68] Y. Nie et al. "Deep Melanoma classification with K-Fold Cross-Validation for Process optimization". In: *2020 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*. 2020, pp. 1–6. DOI: `10.1109/MeMeA49120.2020.9137222`.

[69] Yu Zeng, Kebei Jiang, and Jie Chen. "Automatic Seismic Salt Interpretation with Deep Convolutional Neural Networks". In: *Proceedings of the 2019 3rd International Conference on Information System and Data Mining*. ICISDM 2019. Houston, TX, USA: Association for Computing Machinery, 2019, pp. 16–20. ISBN: 9781450366359. DOI: `10.1145/3325917.3325926`.

[70] Jason Brownlee. *A Gentle Introduction to k-fold Cross-Validation*. 2018. URL: `https://machinelearningmastery.com/k-fold-cross-validation/`.

[71] Anders Krogh and John Hertz. "A Simple Weight Decay Can Improve Generalization". In: *Advances in Neural Information Processing Systems*. Ed. by J. Moody, S. Hanson, and R. P. Lippmann. Vol. 4. Morgan-Kaufmann, 1992. URL: `https://proceedings.neurips.cc/paper/1991/file/8eefcfdf5990e441f0fb6f3fad709e21-Paper.pdf`.

[72] Nitish Srivastava et al. "Dropout: A Simple Way to Prevent Neural Networks from Overfitting". In: *Journal of Machine Learning Research* 15.56 (2014), pp. 1929–1958. URL: `http://jmlr.org/papers/v15/srivastava14a.html`.

[73] Alex Hernandez-Garcia and Peter König. *Data augmentation instead of explicit regularization*. 2020. URL: `https://openreview.net/forum?id=H1eqOnNYDH`.

[74] Jack Scantlebury et al. "Data Set Augmentation Allows Deep Learning-Based Virtual Screening to Better Generalize to Unseen Target Classes and Highlight Important Binding Interactions". In: *Journal of Chemical Information and Modeling* 60.8 (2020). PMID: 32701288, pp. 3722–3730. DOI: `10.1021/acs.jcim.0c00263`.

[75] Leslie Rice, Eric Wong, and Zico Kolter. ”Overfitting in adversarially robust deep learning”. In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by Hal Daumé III and Aarti Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, July 2020, pp. 8093–8104. URL: http://proceedings.mlr.press/v119/rice20a.html.

[76] Jason Brownlee. *A Gentle Introduction to Transfer Learning for Deep Learning*. 2017. URL: https://machinelearningmastery.com/transfer-learning-for-deep-learning/.

[77] Alberto Garcia-Garcia et al. *A Review on Deep Learning Techniques Applied to Semantic Segmentation*. 2017. arXiv: 1704.06857 [cs.CV].

[78] S. S. Varshney, N. Rajpal, and R. Purwar. ”Comparative study of image segmentation techniques and object matching using segmentation”. In: *2009 Proceeding of International Conference on Methods and Models in Computer Science (ICM2CS)*. 2009, pp. 1–6. DOI: 10.1109/ICM2CS.2009.5397985.

[79] Neeraj Sharma and Lalit M Aggarwal. *Automated medical image segmentation techniques.* 2010. DOI: 10.4103/0971-6203.58777.

[80] Mohammad Hesam Hesamian et al. *Deep Learning Techniques for Medical Image Segmentation: Achievements and Challenges*. 2019.

[81] Intisar Rizwan I Haque and Jeremiah Neubert. ”Deep learning approaches to biomedical image segmentation”. In: *Informatics in Medicine Unlocked* 18 (2020), p. 100297. ISSN: 2352-9148. DOI: https://doi.org/10.1016/j.imu.2020.100297.

[82] N. Dhungel, G. Carneiro, and A. P. Bradley. ”Deep structured learning for mass segmentation from mammograms”. In: *2015 IEEE International Conference on Image Processing (ICIP)*. 2015, pp. 2950–2954.

[83] Mugahed A. Al-antari et al. ”A fully integrated computer-aided diagnosis system for digital X-ray mammograms via deep learning detection, segmentation, and classification”. In: *International Journal of Medical Informatics* 117 (2018), pp. 44–54. ISSN: 1386-5056. DOI: https://doi.org/10.1016/j.ijmedinf.2018.06.003.

[84] D. Mishra et al. ”Ultrasound Image Segmentation: A Deeply Supervised Network With Attention to Boundaries”. In: *IEEE Transactions on Biomedical Engineering* 66.6 (2019), pp. 1637–1648. DOI: 10.1109/TBME.2018.2877577.

[85] Boris Sekachev, Nikita Manovich, and Andrey Zhavoronkov. *Computer Vision Annotation Tool*. Oct. 2019. DOI: 10.5281/zenodo.3497106.

[86] Adam Paszke et al. ”PyTorch: An Imperative Style, High-Performance Deep Learning Library”. In: *Advances in Neural Information Processing Systems 32*. Ed. by H. Wallach et al. Curran Associates, Inc., 2019, pp. 8024–8035. URL: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

[87] Karen Simonyan and Andrew Zisserman. *Very Deep Convolutional Networks for Large-Scale Image Recognition*. 2015. arXiv: `1409.1556 [cs.CV]`.

[88] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. *U-Net: Convolutional Networks for Biomedical Image Segmentation*. 2015. arXiv: `1505.04597 [cs.CV]`.

[89] Tsung-Yi Lin et al. *Feature Pyramid Networks for Object Detection*. 2017. arXiv: `1612.03144 [cs.CV]`.

[90] Kaiming He et al. *Deep Residual Learning for Image Recognition*. 2015. arXiv: `1512.03385 [cs.CV]`.

[91] W. R. Crum, O. Camara, and D. L. G. Hill. ”Generalized Overlap Measures for Evaluation and Validation in Medical Image Analysis”. In: *IEEE Transactions on Medical Imaging* 25.11 (2006), pp. 1451–1461. DOI: `10.1109/TMI.2006.880587`.

[92] Wei Qian, Wenqing Sun, and Bin Zheng. ”Improving the efficacy of mammography screening: the potential and challenge of developing new computer-aided detection approaches”. In: *Expert Review of Medical Devices* 12.5 (2015). PMID: 26174018, pp. 497–499. DOI: `10.1586/17434440.2015.1068115`.

[93] F. Pérez-García, R. Sparks, and Sébastien Ourselin. ”TorchIO: a Python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning”. In: *ArXiv* abs/2003.04696 (2020).

[94] Sajid Iqbal et al. ”Deep learning model integrating features and novel classifiers fusion for brain tumor segmentation”. In: *Microscopy Research and Technique* 82 (Apr. 2019). DOI: `10.1002/jemt.23281`.

[95] Zeshan Hussain et al. ”Differential Data Augmentation Techniques for Medical Imaging Classification Tasks”. In: (2017). DOI: `https://doi.org/10.1002/jemt.23281`.

[96] Caglar Senaras et al. ”DeepFocus: Detection of out-of-focus regions in whole slide digital images using deep learning”. In: *PLOS ONE* 13.10 (Oct. 2018), pp. 1–13. DOI: `10.1371/journal.pone.0205387`.

[97] Xiangyu Xu et al. ”Learning to Super-Resolve Blurry Face and Text Images”. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. Oct. 2017.