ILC detection

Applying image processing and deep learning to improve the detection of Invasive Lobular Carcinoma using mammography

Elsemiek Smilde

Technische Universiteit Delft

TU Delft

Erasmus MC

ILC detection

Applying image processing and deep learning to improve the detection of Invasive Lobular Carcinoma using mammography

by

Elsemiek Smilde

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Tuesday January 28, 2021 at 09:30 AM.

Student number:	4366263	
Project duration:	March 1, 2021 – January 28, 2022	
Thesis committee:	Prof. Dr. ir. M.B. van Gijzen,	TU Delft, Numerical Analysis
	Dr. ir. E.G. Rens,	TU Delft, Mathematical Physics
	Dr. M.J. Hooning,	Erasmus MC, Medical Oncology, Cancer Epidemiology group
	Dr. B.A.M. Heemskerk-Gerritsen,	Erasmus MC, Medical Oncology, Cancer Epidemiology group
	Dr. ir. S. Klein,	Erasmus MC, Biomedical Imaging Group Rotterdam

TU Delft

Abstract

Deep learning is a growing field of research and and so is the application of deep learning to the analysis of medical images. Convolutional neural networks are used to diagnose diseases, determine risk of disease development, finding the exact area of abnormalities and and so on. Mammography is an imaging technique, which aims at the early detection of breast cancer. Invasive Lobular Carcinoma (ILC) is a type of breast cancer with properties that make it less visible on mammography. This study compares six models which apply convolutional neural networks to detect breast cancer, on its ability to detect ILC. Furthermore, transfer learning with ILC and healthy images is applied to one of these models to improve the performance on ILC data.

For the evaluated breast cancer detection models, the performance on ILC data is worse than for a dataset which includes all breast cancer types and more healthy images. The model that transfer learning is applied to performs better on ILC data after transfer learning than before, with an increase of 0.09 for the AUC value. Additional analyses of the results show that women with high breast density have a lower chance of getting a correct ILC diagnosis from the model than women with low beast density and this also holds for other types of breast cancer. Lastly, the model outcomes are compared to radiologist reviews, to determine the additional value of models to the routine screening performed by radiologists. Within the images that are labeled as healthy by the radiologists, a model could be applied to detect tumor that have been missed by radiologists. When a specificity of 89% was allowed, 23% of the missed tumors could have been detected by the original GMIC model. In this way, the models used in this study and other deep learning models that are in development now, can contribute to breast cancer detection from mammography, and ILC specifically.

Preface

This thesis is the final project of my MSc Applied Mathematics, and it also represents the end of my student period at the TU Delft, in which I had an amazing time. I started this period as a student of Technology, Policy and Management, but realized after a year what my true calling was in life: mathematics. I am still very happy with this decision and everything that happened afterwards, eventually leading to this moment. In this project I got the possibility to use the skills I have learned over the years in a real application, that also makes a social impact and I am very grateful for that.

I would like to thank everyone that helped me and guided me during this project. I had the luck of being supervised by a group of people from different backgrounds, which enabled me to learn a lot about various fields. I would like to thank Martin and Lisanne for their feedback from a mathematical and analytical perspective, Maartje and Annette for their views from the medical side and Stefan for all your knowledge on deep learning and for helping me guide through the (for me very confusing at the beginning) image storage at the Erasmus MC. I was very lucky to be a part of research groups at the Erasmus MC and have the opportunity to get to know other people and research subjects and have conversations to gain new ideas, get inspired or just to have fun. Thank you all for letting me feel so welcome.

Elsemiek Smilde
Rotterdam, 24 January 2022

Contents

Abbreviations

AUC Area Under the Curve. 20

BCE Binary Cross Entropy. 30

CAD Computer-aided detection. 7

CNN Convolutional Neural Network. 15

CSAW Cohort of Screen-Aged Women. 24

EMC Erasmus Medical Centre. 23

IDC Invasive Ductal Carcinoma. 2

ILC Invasive Lobular Carcinoma. 2

KI Karolinska Institute. 23

MRI Magnetic Resonance Imaging. 1

ReLu Rectified Linear Unit. 16

ROC Receiver Operating Characteristic. 19

List of Figures

List of Tables

1 Introduction

The first four sections of this chapter serve as an introduction to the topics of breast cancer and mammography, and afterwards the research aim, research questions and outline of the report will be stated.

1.1. Breast cancer

In the Netherlands, approximately 15,000 women are diagnosed with breast cancer yearly (IKNL, n.d.). Breast cancer may be detected in several ways. In case of symptomatic breast cancer, women will experience a lump or abnormal shape in their breast or irregular pain. In that case, women can undergo a breast exam, where a doctor will check both breasts and lymph nodes in their armpit, feeling for any lumps or other abnormalities. Additionally, mammograms and biopsies are used for the diagnosis. Another way of early detection of breast cancer is The National Breast Cancer Screening Programme. This programme is designed for women between 50 and 75 years of age, and specifically aims to detect asymptomatic breast cancer, which is the case when women do not experience symptoms. Once every 2 years, women in this age group are invited for a mammogram. Other used imaging techniques to detect breast cancer are Breast Ultrasound and Breast Magnetic Resonance Imaging (MRI). After an abnormality has been observed, a biopsy can be taken, which is the removal of a sample of breast cells for pathological examination, using a microscope. A biopsy is the only definitive way to confirm a diagnosis of breast cancer.

1.2. Mammography

This research is focused on mammographic images. Mammography is the process of using low-energy x-rays to examine the breast for diagnosis and screening. The goal of mammography screening is the early detection of breast cancer, typically through detection of characteristic masses or microcalcifications. During a mammogram study, the breast of a woman is placed between two parallel surfaces and compressed between them. A small beam of x-rays are produced by a machine, sent through the breast tissue and captured on the other side by a detector. Nowadays, this detector often is a solid-state detector and produces a digital image, by sending electronic signals to a computer. Another detector that was used more before is a photographic film plate. The images that are produced by mammography are called mammograms.

These mammograms show differences in brightness in the different tissue types. Tissues with low density, such as fat, appear darker on the images, while tissues with high density, such as connective and glandular tissue or tumors, appear lighter.

A mammography exam consists of four images. For each breast two images, resulting in four images. The images are made from two different angles. One angle is the Cranial-Caudal (CC) view, which is a view from above. The other most used view is the Mediolateral-Oblique (MLO) view, which is a view from the side at an angle. Consequently, the four images resulting from a mammography screening are *R CC*, *L CC*, *R MLO* and *L MLO*, where *L* indicates the left breast, and *R* indicates the right breast. Sometimes, extra views may be taken when, after first evaluation, the physician is concerned about a specific area of the breast.

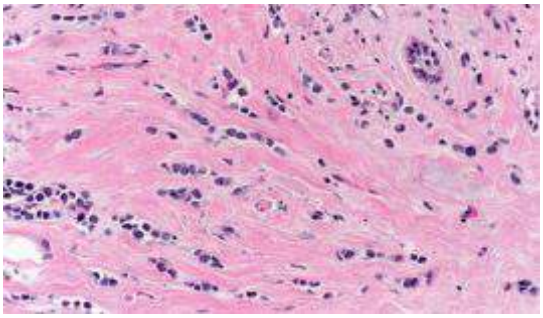
The way how mammograms are interpreted are different across the world. In the US, one radiologist evaluates the mammograms and performs the detection task (single-radiologist reading), whereas in Europe and Asia usually two radiologists do this together (double reading) (Mello-thoms, 2020). In the case of single-radiologist reading, false-negative rates are between 10-30% (Wadhwa et al., 2016). This results in a sensitivity of 70-90%. Also, around 49% of all women annually scanned with mammography will receive at least one false-positive test result (Wadhwa et al., 2016). According to The European Guidelines for quality assurance in breast cancer screening and diagnosis (Perry et al., 2008), sensitivity increases by 5-15% in the case of double reading. Other studies on the sensitivity of double reading report values between 72% and 86,9% (Kavanagh et al., 2000; Lehman et al., 2017; Lehman et al., 2015; von Euler-Chelpin et al., 2018) for test populations, comparable to a screening program.

There are several possible causes for breast cancer to go undetected on mammography (false-negatives) (Wadhwa et al., 2016). These include the following:

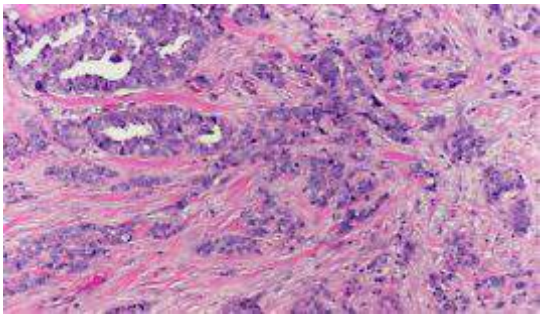
- Asymmetry, which is defined as a planar finding that is seen on only 1 of the 2 standard mammographic views and that lacks convex margins.
- Perception errors, when a finding is included in the field of view, but not detected by the radiologist. Reasons for this can be a lack of experience or attention,
- Misinterpretation, when a finding is interpreted as benign, but it is actually malignant.
- Poor technique and limitations of the screening methods.
- High breast density, where the dense tissue can obscure underlying small tumors.
- Low visibility of the tumor, e.g. in the case of Invasive Lobular Cancer, which is discussed further below in section 1.3.

1.3. Invasive Lobular Carcinoma

The type of breast cancer that is investigated in this study is Invasive Lobular Carcinoma (ILC). ILC is the second most common subtype of breast cancer after Invasive Ductal Carcinoma (IDC), accounting for up to 15% of all breast cancer cases (Christgen et al., 2016). IDC is a form of breast cancer that has its origin in a milk duct and is invading breast tissue outside of the duct. ILC is also an invasive cancer, but it began growing in milk-producing glands (lobules). ILC presents a number of unique challenges for diagnosticians and treating physicians. Compared to IDC, ILC is more often not palpable. Furthermore, ILC presents more frequently in several locations within one breast (multifocal/multicentric) or in both breasts (bilateral) (Christgen et al., 2016; Desmedt et al., 2017). During the diagnosis of ILC, generally there are more involved lymph nodes than for IDC, which can be a sign of tumor in a later stage. This suggests that ILC is often detected later than IDC. In terms of biological characteristics, oncologists look at the presence of hormone receptor proteins, which influence the growth of the tumor. The presence or absence of these proteins influence the decision of treatment options. ILC shows more frequently expression of estrogen receptor (ER) and progesterone recepter(PR), and less amplification of human epidermal growth factor receptor 2 (HER2) (Chen et al., 2017; Christgen et al., 2016; Desmedt et al., 2017; Guiu et al., 2014). Lastly, there is usually a lack of E-cadherine expression. Possibly due to this last characteristic, tumor cells are often small, round and discohesive, since E-cadherine is a cell-cell adhesion molucule. Therefore, the cells lie in loose columns rather than in a lump. This is shown on cell-level and compared to IDC in Figure 1.1.



(a) Malignant cells of Invasive Lobular Carcinoma form lines



(b) Malignant cells of Invasive Ductal Carcinoma form ducts or tubules

Figure 1.1: Comparison between IDC (a) and ILC (b) tumorous cells (John Hopkins University, 2021)

Some of these features of ILC complicate accurate imaging of the tumor, since the tumor usually does not present as a clearly defined mass (Desmedt et al., 2017). This complicates detection with imaging techniques. During the population screening, ILC is quite often missed. Even in retrospect, at least 20% of cases are not visible at mammography. As a consequence, ILC is often not detected at an early stage. Delayed breast cancer diagnosis may cause several problems, including larger tumor size, more lymph node involvement and more distant metastases. As a result, delayed detection may result in a worse prognosis.

In previous studies, imaging techniques to detect invasive lobular carcinoma have been compared. Brem et al. (2009) compared the sensitivity of mammography, sonography, MRI, and breast-specific gamma imaging (BSGI) in the detection of ILC. They compared, for 28 biopsy-proven invasive lobular carcinoma of 26 women, the results between different imaging modalities when interpreted by radiologists. Even though this is a small number, the results give some insights in the abilities of the imaging methods. BSGI had the highest sensitivity for the detection of invasive lobular carcinoma with a sensitivity of 93%, whereas mammography, sonography, and MRI showed sensitivities of 79%, 68%, and 83%, respectively. Other research shows a sensitivity between 57% and 79% for mammography in detecting ILC (Johnson et al., 2015). To compare this with other types of breast cancer, Kerlikowske et al. (1996) found that in general, sensitivity in detection of invasive breast cancers from mammography lies between 58% and 93%, which is on average higher than the sensitivity of ILC detection. In Figure 1.2 an example is given of mammograms with ILC.

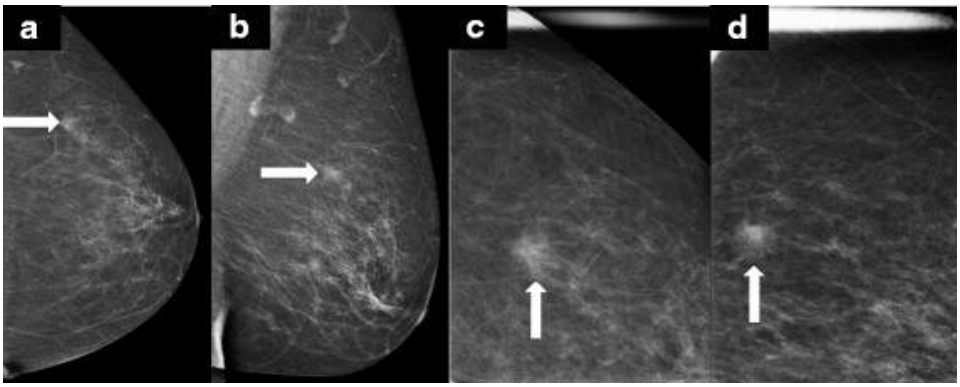


Figure 1.2: Invasive lobular carcinoma presenting as a mass on mammography (Johnson et al., 2015)

Hilleren et al. (1991) analysed 137 mammograms from ILC cases to find common abnormalities in the images. The following abnormalities were detected in the images, together with the percentage of images with the abnormality found:

- **Spiculated opacity (53%)**: an area can be observed that is lighter on the mammogram than the rest of the breast tissue
- **Architectural distortion (16%)**: a deviation or retraction of the normal pattern of the breast tissue
- **Poorly defined opacity (7%)**: an area where some lighter sections can be found, but less well defined than in the case of spiculated opacity
- **Normal or benign findings (16%)**: no indications of malignancies
- **Parenchymal assymetry (4%)**: differences in the tissue between the two breasts

Furthermore, the craniocaudal view of the breast showed the abnormalities more frequently than the oblique or lateral views. The craniocaudal view is shown in the first image (a) of Figure 1.2, and is a view in the horizontal direction through the breast tissue. The mediolateral oblique view is shown in (b) of Figure 1.2 and is a view in the diagonal direction through the breast tissue.

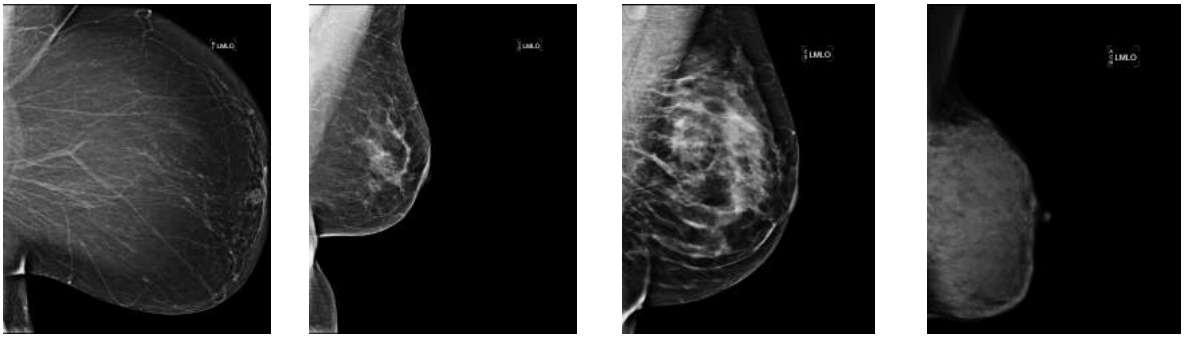
1.4. Breast density and issues in the detection

As stated in section 1.2 high breast density can be a reason for a breast cancer to go undetected in mammography. The dense tissue can cover up the cancerous tissue on the images. Women with extremely dense

breast also have an increased risk of breast cancer (Boyd et al., 2007), which increases the impact of the failed detection.

There are four density categories defined, of which example mammograms are shown in Figure 1.3. The categories are the following:

1. Almost entirely fatty
2. Scattered fibroglandular density
3. Heterogeneously dense
4. Extremely dense



(a) Density category 1 (b) Density category 2 (c) Density category 3 (d) Density category 4

Figure 1.3: Example mammograms for each density category, from the CSAW dataset

A study by Wanders et al. (2017) looked into the interval cancer rate, false-positive rate and screening sensitivity for different breast densities. An interval breast cancer occurs, when breast cancer is diagnosed not during a screening mammogram, but during the time in between two screening mammograms. The results are shown in Table 1.1. These results show that mammographic images of breasts with a high density category, are more often wrongly diagnosed compared to less dense breasts. Consequently, women with high breast density are more likely to have an interval cancer.

	Breast density			
	1	2	3	4
Interval cancer rate (%)	0.7	1.9	2.9	4.4
False positive rate (%)	11.2	15.1	18.2	23.8
Screening sensitivity (%)	87.7	77.6	69.5	61.0

Table 1.1: Detection results for different breast density categories (Wanders et al., 2017)

A study by Bakker et al. (2019) looked into the use of magnetic resonance imaging (MRI) for women with high breast density as an additional screening method, to improve breast cancer detection. They found significantly fewer interval cancers in a group that underwent this MRI screening, compared to a control group that only underwent mammography screening. Both groups had a negative result from the mammography screening, indicating no tumor was shown on mammography. The MRI-screened group had 2.5 interval cancers per 1000 screenings, whereas the control group had 5.0 interval cancers per 1000 screenings. Focusing on ILC specifically, 9 out of 11 cancers (82%) that developed during the study period were detected through MRI. For IDC, 35 out of 37 (95%) developed cancers were detected through MRI. This shows that MRI could highly improve the detection of ILC, but ILC is still more often missed than IDC.

1.5. Research aim and questions

The goal of this project is to improve the detection of ILC with mammography, thereby providing detection of ILC in an earlier stage, and thus improving prognosis of ILC patients. This is done by applying various methods on the collected mammograms. The main research question is:

Can image processing techniques and deep learning models contribute to the earlier detection of Invasive Lobular Carcinoma from mammograms?

The study covers two sub topics. The first focus is on image processing techniques like described in chapter 2 and the second focus is on deep learning models, like described in chapter 3. Lastly, the influence of breast density on these results should be considered. Therefore, the following subquestions can be formulated:

1. Is it possible to improve the visibility of ILC on mammograms using image processing techniques?
2. Is it possible to perform automatic detection of ILC using deep learning frameworks?
   (a) How do current existing deep learning models for breast cancer detection perform on detecting ILC?
   (b) Is it possible to use transfer learning to improve the ability of a model to detect ILC?
3. What are the results for ILC detection for different breast density categories?
4. Could the developed method add value to the detection of ILC, done by breast radiologists?

1.6. Project phases

For this project, six phases are distinguished:

1. Testing image processing techniques in an explorative way on 20 mammograms
2. Validate these methods on a new set of approximately 50 mammograms, which contain both visible tumors and poorly detectable ones.
3. Testing multiple existing deep learning models on the available datasets
4. Selecting a model for transfer learning on ILC data, performing transfer learning, and validation
5. Finding a method to automatically assess breast density and use these results to find connections between breast density and detection accuracy
6. Lastly, determining the effectiveness of the method in detecting ILC in an earlier stage, by comparing the model performance to radiologist reviews

The datasets that are used are described in chapter 4, and the explanation of the experiments in chapter 5 also state the used images for each experiment.

1.7. Outline of report

The chapters in this report are structured as follows. Chapter 2 describes the available literature on image processing and chapter 3 goes deeper into deep learning methods for breast cancer detection from mammography. In chapter 4 the two different data-sets that are used in this research are described. Chapter 5 describes the methods, used in the experiments and how they are applied. These methods are split into image processing techniques and deep learning. Furthermore, the metrics used for evaluation are given and explained. The results are presented in chapter 6. A discussion of the results is presented in chapter 7, and this also includes limitations, comparisons to previous research and recommendations for future research. Finally, a conclusion is presented in chapter 8.

1.8. Review of Medical Ethics Committee Erasmus MC

The Daily Board of the Ethics Committee Erasmus MC of Rotterdam, The Netherlands, has reviewed the research proposal. Based on this, they decided that the rules laid down in the Medical Research Involving Human Subjects Act (WMO) do not apply to this study. Therefore, provided that the images are anonymised, the study can be conducted.

2

Image processing techniques

Image processing is the technique in which an image is used as input, and this is translated into another image as output. Conventional image processing techniques include all methods, that do not depend on training data, in contrast to the deep learning techniques discussed in chapter 3. This chapter explains the following methods: contrast enhancing methods, edge detection and enhancement, segmentation and wavelet transform. These methods are chosen, since they have been applied in previous studies to mammography and their application has been shown to be effective in some cases. Some of these methods have been used in clinical settings in the past decades, especially in the US (Sechopoulos & Mann, 2020). These systems are called Computer-aided detection (CAD) systems, which use image processing techniques and pattern recognition methods for detection and classification of irregularities in mammograms. However, this resulted in many false positives and more work for radiologists. Lehman et al. (2015) compared the diagnostic accuracy of breast cancer detection from mammography for radiologists using CAD to radiologists not using CAD and found no significant difference. Therefore, the conventional image processing and mass detection techniques are now mostly used in retrospective research instead of clinical applications. There are no results yet on the efficiency of these methods on ILC specifically. Therefore, an aspect of the goal of this study is to evaluate the results of the methods on images containing ILC. In the following sections the notation for a not-processed image will be $I(x, y)$ with $(x, y) \in \mathbb{N}^2$ the pixel coordinates. The image that is used to demonstrate some of the methods, is obtained from the mini-MIAS database of mammograms (Suckling et al., 2015).

2.1. Contrast enhancing methods

The contrast of an image reflects the differences between the pixel values in this image. A histogram is a discrete function $h(r_k) = n_k$, where r_k is the kth intensity value and n_k is the number of pixels in the image with intensity r_k (Gonzalez & Woods, 2009). A histogram of an image forms the basis of numerous spatial domain processing techniques. In this part the focus is on intensity transformations using histograms, the first one being contrast enhancement or contrast stretching. Contrast stretching is a method which increases the differences between pixel values, by using the whole available value range to display the pixel values in the image. Let I be the image and $(x, y) \in \mathbb{N}^2$ the pixel coordinates. Let X be the maximum x-coordinate and Y be the maximum y-coordinate. The transformation is given by (Young et al., 2007):

$$b(x,y) = \left(2^B - 1\right) * \frac{I(x,y) - \min_{y\in[0,Y],x\in[0,X]}(I(x,y))}{\max_{y\in[0,Y],x\in[0,X]}(I(x,y)) - \min_{y\in[0,Y],x\in[0,X]}(I(x,y))} \tag{2.1}$$

in which $2^B - 1$ is the highest possible pixel value. Since this formula can be sensitive to outliers there is a more general version formulated, where the highest and lowest values are not considered for the stretching and therefore the values in between this range are stretched even more. This transformation is given by:

$$b(x,y) = \begin{cases} 0 & I(x,y) \leq p_{low}\% \\ \left(2^B - 1\right) * \frac{I(x,y) - p_{low}\%}{p_{high}\% - p_{low}\%} & p_{low}\% < I(x,y) < p_{high}\% \\ \left(2^B - 1\right) & I(x,y) \geq p_{high}\% \end{cases} \tag{2.2}$$

Another intensity transformation which depends on the shape of the histogram is histogram equalization. In this method, a transformation takes place which changes the histogram in the direction of a uniform distribution. This results in an image with the pixels equally divided over all pixelvalues. When an image has a large area with dark pixels for example, the differences in the dark shades will become more visible. The transformation is given by the following, with r the original pixel value, and $f(r)$ the new pixel value:

$$f(r) = \left(2^B - 1\right) * \int_0^r p_a(w)dw \tag{2.3}$$

with $p_a(w)$ being the probability density function of the image I. A variation to histogram equalization is histogram matching. During histogram equalization, the histogram is transformed into a uniform distribution. In histogram matching, the histogram is transformed into another distribution, which can be described by a function or derived from another image. This could be useful when the desired histogram is known beforehand or obtained from another image.

The last intensity transformation to discuss is local or adaptive equalization. The previous techniques are all applied globally to all pixels, whereas local equalization looks at the surrounding pixels of a pixel to determine the new value. In this way, details over small areas can be enhanced, while they would fade away in a global approach. To see the effects of these methods on mammograms, in Figure 2.1 the results are shown for the described methods for one example mammogram, including a graph showing the histogram and cumulative density function. From this images some properties of the methods can be observed. Firstly, histogram equalization and adaptive equalization both make the outer area of the breast more visible, and give a better view of the thin fat tissue surrounding the dense tissue. Secondly, in all three methods the bright area - which is the tumor - is displayed more clearly than in the 'low contrast image'. Thirdly, adaptive equalization shows much more contrast in the breast tissue than the other methods and the original image.

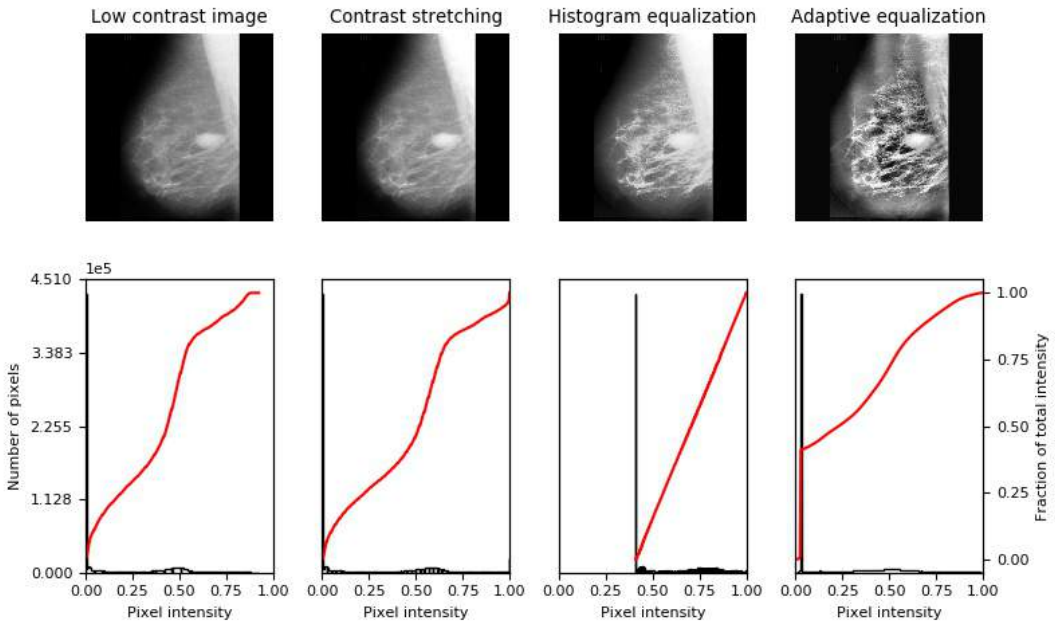


Figure 2.1: Results of intensity transformation methods on test mammogram

2.2. Edge detection and enhancement

The boundary and location of a tumor is of interest to clinicians during diagnosis. Therefore, the shape and location of edges are important properties, since they define boundaries. In this section, methods to locate shapes and edges of objects are explored. Also, a method to make edges more visible in images is explained.

2.2.1. Edge detection based on gradient

To gain insights in differences between neighbouring pixels, the derivative is taken into account. The derivative of an image can be taken in two directions, the horizontal derivative $\mathbf{h_x}$ and the vertical derivative $\mathbf{h_y}$ (Young et al., 2007). A gradient filter generates a vector derivative, which is defined as

$$\nabla f = \text{grad}(f) = \begin{bmatrix} g_x \\ g_y \end{bmatrix} = \begin{bmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \end{bmatrix} \tag{2.4}$$

This vector points to the direction of the greatest rate of change of f at the location of (x, y). The magnitude and direction of the gradient is

$$|\nabla f| = \sqrt{g_x^2 + g_y^2} \tag{2.5}$$

$$\psi\left(\nabla f\right) = \tan^{-1}\left(\frac{g_y}{g_x}\right) \tag{2.6}$$

There are multiple functions to determine the derivative or gradient of an image. Basic derivative filters are the following:

$$\left[\mathbf{h_x}\right] = \left[\mathbf{h_y}\right]^t = \begin{bmatrix} 1 & -1 \end{bmatrix} \tag{2.7}$$

$$\left[\mathbf{h_x}\right] = \left[\mathbf{h_y}\right]^t = \begin{bmatrix} 1 & 0 & -1 \end{bmatrix} \tag{2.8}$$

Then there are the Prewitt gradient filters:

$$\left[\mathbf{h_x}\right] = \frac{1}{3}\begin{bmatrix} 1 & 0 & -1 \\ 1 & 0 & -1 \\ 1 & 0 & -1 \end{bmatrix} \tag{2.9}$$

$$\left[\mathbf{h_y}\right] = \frac{1}{3}\begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ -1 & -1 & -1 \end{bmatrix} \tag{2.10}$$

and the Sobel gradient filters:

$$\left[\mathbf{h_x}\right] = \frac{1}{4}\begin{bmatrix} 1 & 0 & -1 \\ 2 & 0 & -2 \\ 1 & 0 & -1 \end{bmatrix} \tag{2.11}$$

$$\left[\mathbf{h_y}\right] = \frac{1}{4}\begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix} \tag{2.12}$$

The results of applying the Sobel and Prewitt filters on the test image are shown in Figure 2.2.

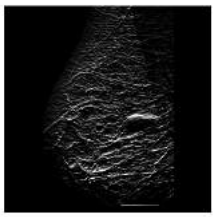
Using the gradient, like described above, is the first way to detect edges. After the application of a gradient method, a threshold is used, which results in the pixel values that have a high enough gradient and are therefore appointed as edges (Young et al., 2007). This method works well when the noise level in the image is low. Otherwise, smoothing techniques should be applied before calculation of the gradient.

2.2.2. Unsharp masking

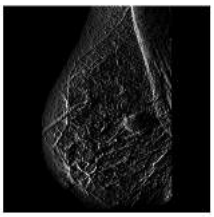
A way to enhance the edges in an image is unsharp masking. In this technique, a low-pass version of the image is subtracted from the image. The following equation shows this:

$$b(x, y) = I(x, y) - \left(k * \nabla^2 I(x, y)\right) \tag{2.13}$$

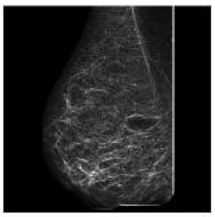
where k is the amount of times the blurred version of the image is subtracted, and a radius is the σ-parameter in the Gaussian filter ∇. The goal of applying unsharp masking on mammograms, is to enhance the edges of mass lesions (Pisano et al., 2000). In Figure 2.3 the results are shown for different values of k. From these images it can be seen that unsharp masking indeed enhances the edges of the mass. It can be observed that especially in the image with radius = 5, the small edges and lines become more visible. In the image with radius = 20, bigger structures become more clear, but small details are less visible.



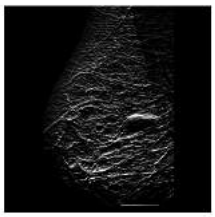
(a) Prewitt, horizontal edges



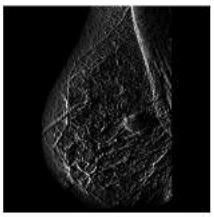
(b) Prewitt, vertical edges



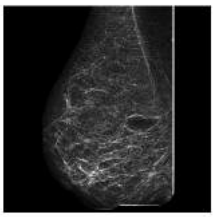
(c) Prewitt, edge magnitude



(d) Sobel, horizontal edges



(e) Sobel, vertical edges



(f) Sobel, edge magnitude

Figure 2.2: Results of gradient methods

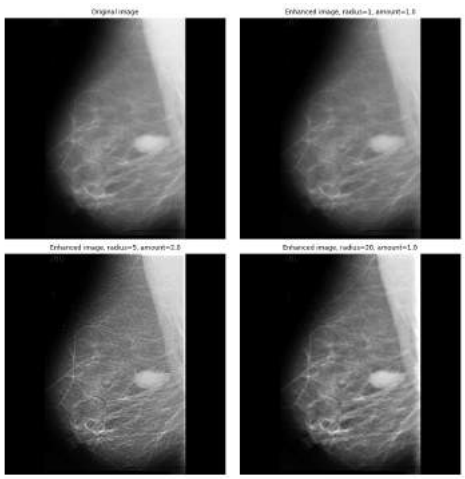


Figure 2.3: Unsharp masking

2.3. Thresholding & segmentation

In medical imaging, edge detection is mostly used for segmentation, in which an image is partitioned into sets of pixels to locate certain objects or boundaries. Segmentation could be used for different purposes in mammograms. Firstly, the breast could be segmented from the background. Secondly, the dense tissue could be segmented from the fat breast tissue. Lastly, segmentation could be used to determine the area of the tumor tissue, which could be very useful in this project. Methods to perform segmentation include thresholding, region growing, edge-based segmentation, watershed segmentation and active contours (Hemalatha et al., 2018).

2.3.1. Thresholding

Probably the most simple method for segmentation is thresholding. The idea behind thresholding is a choice of *brightness threshold* θ and use this to divide the image into pixels, which intensity is higher than θ and

pixels, which intensity is lower than θ (Young et al., 2007):

$$\textbf{If } I_{thresholding}(x,y) \geq \theta \qquad I(x,y) = 1 \tag{2.14}$$

$$\textbf{Else} \qquad I(x,y) = 0 \tag{2.15}$$

In the same way, multiple thresholds can be defined, to split the pixels into more groups. For mammograms, the background consist of very dark pixels. Because of this, thresholding can be used to segment the image into tissue and background. The result of this are shown in Figure 2.4 for the test image. Furthermore, thresholding can be used to determine which pixels belong to dense tissue in a image, since these tissues result in lighter areas and higher pixel values. Figure 2.5 shows results of threshold segmentation for breast tissue for different type of breasts, where 150 is used as thresholds value between dense and non-dense tissue. From these images, it is clear that there exist differences between patients in the amount of dense tissue in the breast. Especially in breasts with a lot of dense tissue, this complicates the detection of tumors. A possibility to solve this, could be to use different thresholds for different types of breasts. The decision of a threshold can also be done using the histogram of the image. There are multiple ways to do this, including the isodata algorithm, the background-symmetry algorithm and the triangle algorithm (Young et al., 2007). Moreover, instead of one threshold for the entire image, a local or adaptive threshold can be applied. This threshold value is calculated for smaller regions, using pixels close to the pixel of interest.

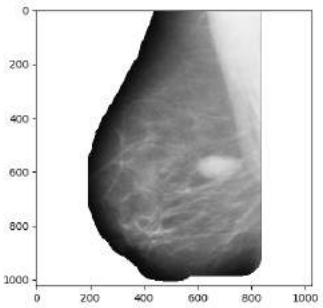
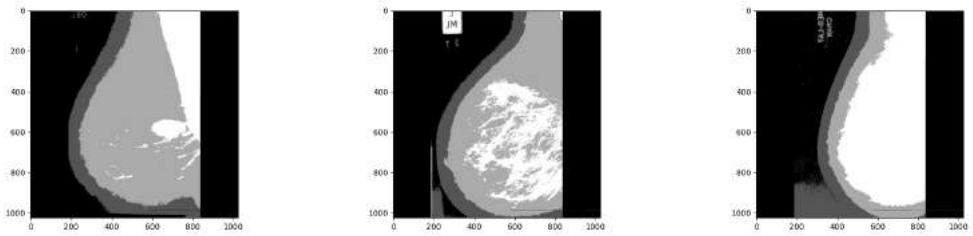


Figure 2.4: Thresholding to remove background



(a) Low density breast tissue (b) Medium density breast tissue (c) High density breast tissue

Figure 2.5: Thresholding segmentation results for different density of breast tissue using one threshold

Kom et al. describe a method for mass detection from mammograms based on thresholding (Kom et al., 2007).

2.3.2. Segmentation

Other techniques to segment objects in images that have been applied to mammograms include watershed segmentation (Preim & Botha, 2014; Young et al., 2007), edge-based segmentation (Zhang et al., 2010), region growing (Alamin et al., 2016; Berber et al., 2013; Gonzalez & Woods, 2009; Preim & Botha, 2014), level sets (J. Liu et al., 2011; Shi et al., 2008) and active contour models (Ferrari et al., 2004; Hemalatha et al., 2018; Preim & Botha, 2014).

2.4. Wavelet transform

The detection of microcalcification in previous research is mainly done using wavelet-based algorithms (Grgic et al., 2009). The goal of these decomposition methods is to find and separate clusters of microcalcifications from the surrounding tissue. This is done by looking at high frequency regions in the images. After a wavelet transform, low-frequency components can be removed and the detection of high frequency areas starts. There exist variations in the decomposition and detection techniques, but the set-up of these methods is comparable. A wavelet transform constructs a time-frequency representation of an image. Furthermore, a wavelet transformation gives the location of the image where the frequency is detected. The 1D Wavelet function Φ is defined as:

$$\Phi_{s,\tau}(t) = \frac{1}{\sqrt{s}}\Phi\left(\frac{t-\tau}{s}\right) \tag{2.16}$$

with $s, t \in \mathbf{R}$ and $s \neq 0$ and

$$\int_{-\infty}^{\infty} \Phi(t)\,dt = 0 \tag{2.17}$$

A wavelet transform in images decomposed an image into four subimages (Strickland & Hahn, 1996):

1. LL, which is formed by low-pass filtering in both directions and therefore shows the low-frequency details in an image. This results in a smoothed image.
2. LH, which is formed by low-pass filtering in the rows and high-pass filtering in the columns, which results in mostly the horizontal details in the image
3. HL, which is formed by low-pass filtering in the columns and high-pass filtering in the rows, which results in mostly the vertical details in the image
4. HH, which shows mostly the diagonal details

For the test-image, the results of the subimages are shown in Figure 2.6. The method to use Wavelet transform for detecting microcalcification is as follows (Strickland & Hahn, 1996). Usually, the highest pixels in HH and LH+HL are detected, weighted and dilated, after which the inverse Wavelet transform is used. Since this procedure enhances the visibility of the microcalcifications, thresholding can be applied afterwards to segment the microcalcifications. An issue with this method is the high rate of false positives, when the method detects an area where no microcalcification is present.

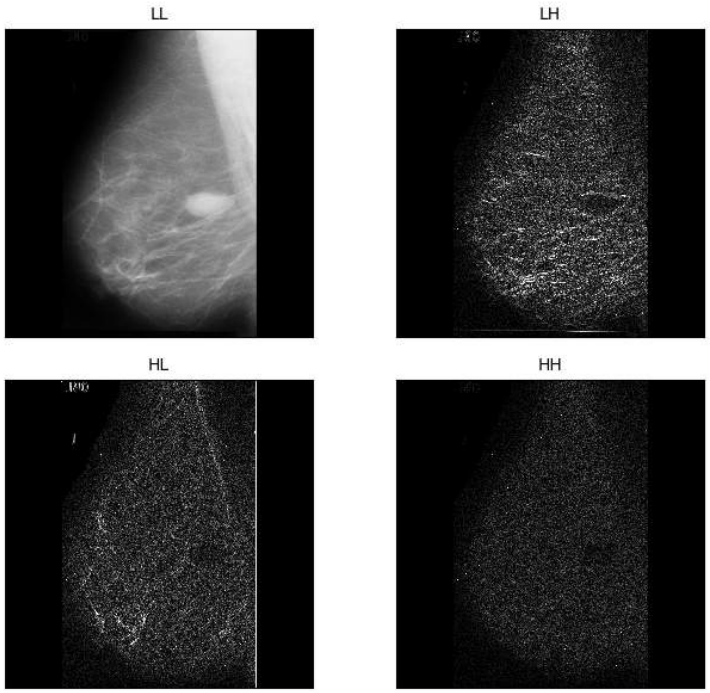


Figure 2.6: Wavelet transform

3

Deep learning

Another recent field of research, considering the detection of breast cancer from mammograms focuses on deep learning. The difference between the conventional image processing methods in the previous chapter and the ones using deep learning, is that deep learning methods use data and learn how to distinguish classes based on this themselves, while for the conventional methods this should be predefined by the developer of the method. The application of deep learning in mammography is a growing field (Sechopoulos & Mann, 2020). Recently, mostly deep learning methods for digital mammograms are being developed. There are some applications of deep learning models in clinical practice. However, there is still a gap between the accuracy that can be achieved with these models and the implementation and use by radiologists (Masud et al., 2019). In general, more research is needed on how these models can contribute to the work of radiologists and their perception on them. So far, there are no studies that focus on the efficiency of the algorithms on the detection of ILC specifically.

3.1. Convolutional Neural Networks

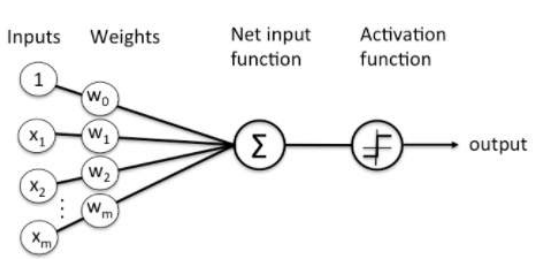
Deep learning is a subfield of machine learning, in which models require a big labeled dataset to be trained. Deep learning models have existed for quite some time, but only recently they became more present in research and other applications due to the growth in computing power, the decrease in hardware cost, open source algorithms and increase in data availability (Abdelhafiz et al., 2019). Neural networks make up the backbone of deep learning. Neural networks consist of a number of connected nodes, called neurons, that are put into layers. A node combines input with weights, sums this up and creates output through an activation function, which determines whether and how the output should be sent through the network. In Figure 3.1a a diagram is shown for what a node might look like. A layer consists of a combination of nodes, that receive input from the previous layer and generate output for the next layer, like shown in Figure 3.1b. Deep learning refers to a high amount of layers stacked together to form the model.

Convolutional Neural Networks are a class of neural networks that are mostly used in image processing (Abdelhafiz et al., 2019). Components of convolutional neural networks are the following:

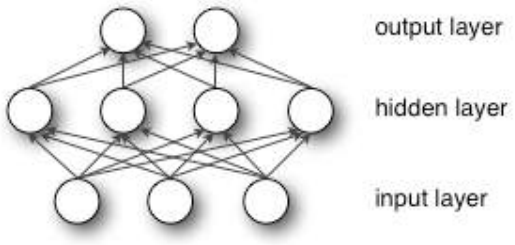
Convolutional layers

These layers consist of a set of filters that are applied on the input data, creating an output - or activation - feature map of that filter using a kernel. Kernels are matrices, which are designed to respond to patterns in the input data. Examples of this are kernels to find horizontal, vertical or diagonal edges. The name 'convolutional' stems from the convolution operation, where the kernel slides over the input image and saves the result at every position. For the first image, the image is the input for this operation, in further layers, the output from previous layers is taken. The convolution operation is defined as follows. Let $I \in \mathbb{R}^{KxL}$ be the input, $K \in \mathbb{R}^{MxN}$ the kernel and S the ouput feature map. S can be calculated as:

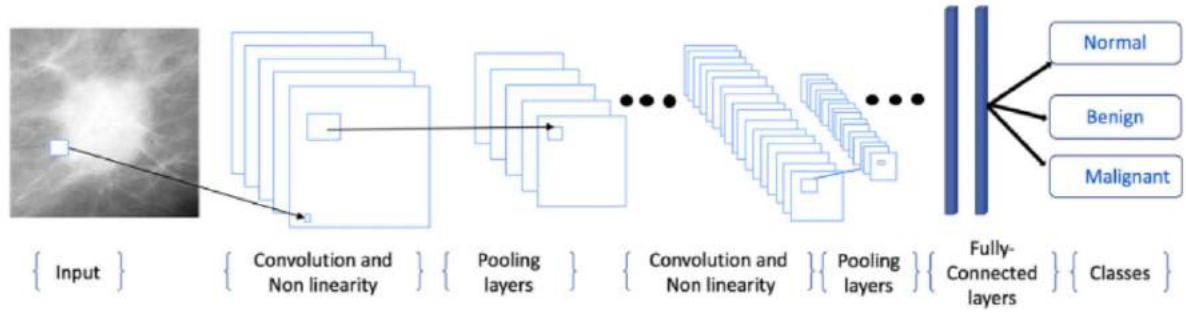
$$S(i,j) \equiv (K * I)(i,j) = \sum_{m}^{M} \sum_{n}^{N} I(i+m, j+n) K(m,n) \tag{3.1}$$



(a) A node in a neural network
URL: https://wiki.pathmind.com/neural-network



(b) Layers of nodes in a neural network
URL: https://wiki.pathmind.com/neural-network



(c) A basic architecture of a convolutional neural network (Abdelhafiz et al., 2019)

Figure 3.1: Elements of a convolutional neural network

Nonlinear layers
These layers are also called activation layers, and used to enable the model to learn non-linear mappings as well. An example is the ReLu (rectified linear unit) layer, which removes negative values from an activation map by setting them to zero, as follows:

$$f(x_{ij}) = \begin{cases} 0 \text{ for } x \leq 0, \\ x \text{ for } x > 0. \end{cases} \tag{3.2}$$

Pooling layers
These layers are used for non-linear down-sampling. Down-sampling means that the output from the previous layer is made smaller, by summarizing some numbers in a single statistic. Similar as in the convolutional layers, a kernel slides over the input image and determines the outcome. The pooling method that is used most in convolutional neural networks, is max-pooling, in which the maximum value in the window of the input map is taken and put into the output map.

Fully connected layers
In these layers, that are placed in the end of the network, there exist connections between all activations in the previous layer and all nodes in the fully connected layer.

Final layer
The final layer usually has a size the same as the number of outputs that the model needs. Finally, an end activation function is needed to determine the result of the model, for example a binary classification or a probability. A sigmoid function or a softmax function are often used for this purpose. The sigmoid function results in a value between 0 and 1 and is defined as follows:

$$f(x_i) = \frac{1}{1 + e^{-x_i}} \tag{3.3}$$

The softmax function is often used when there are multiple classes possible as output for the model. The definition is:

$$f(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{c} e^{x_j}} \tag{3.4}$$

A loss function
A loss function defines how differences between predicted output values and true values are penalized. For models that aim to do binary classification, cross-entropy loss is the most commonly used method. It is defined as follows: Let m be the number of examples in the dataset: $\{X^{(i)}\}_{i=1}^{m} \in \mathbb{R}^{V\text{x}W}$ with true labels $\{Y^{(i)}\}_{i=1}^{m} \in \mathbb{R}^{V\text{x}W}$ and the predicted labels $\{\hat{P}(X^{(i)};\Theta)\}_{i=1}^{m} \in \mathbb{R}^{V\text{x}W}$, with Θ the weights in the network, the cross-entropy loss in then defined as:

$$L(\Theta) = -\frac{1}{m}\sum_{i=1}^{m}\sum_{k=1}^{K} Y_k^{(i)} \log\left(\hat{P}_k(X^{(i)};\Theta)\right) \tag{3.5}$$

For binary classification, the cross-entropy loss is therefore defined as:

$$L(\Theta) = -\frac{1}{m}\sum_{i=1}^{m} Y^{(i)} \log\left(\hat{P}(X^{(i)};\Theta)\right) + (1-Y^{(i)}) \log\left(1-\hat{P}(X^{(i)};\Theta)\right) \tag{3.6}$$

In Figure 3.1c a basic architecture of a convolutional neural network is shown. The output of the model is usually a prediction for a classification, *normal*, *benign* or *malignant* in the figure or a segmentation. The most popular convolutional neural networks are ALex-Net, ZF-Net, GoogleLeNet, VGG-Net and ResNet (Abdelhafiz et al., 2019), which all have a typical architecture and order of layers.

3.2. Deep learning in mammography

First, classical machine learning models were used to detect breast tumors and determine the malignancy of them. The method used for this in most researches is composed of multiple steps (Dogra et al., 2019; Jiménez-Gaona et al., 2020). The first step is preprocessing of the images, in which the image quality and contrast is improved and noise, background pixels and pectoral muscle is removed. The second step is to find the outlining of the region where something might be found, the region of interest (ROI). Then, feature extraction is used to gather information about the ROI in scalars, feature selection to reduce the dimension and then a label is produced based on a machine learning classifier. More recently, convolution neural networks are being developed the most, which remove the step of feature selection, since this is built into the structure. In Table 3.1 an overview is given of studies that use this approach, together with the used architecture, database and application of the model. This overview is created with the help of the review paper from Wong et al. (2020), the review paper from Jiménez-Gaona et al. (2020) and by looking for other studies on this subject. Only recent studies concerning the application of CNNs for mammography are taken into account.

3.2.1. Applications of CNNs in mammograms

As seen in Table 3.1, the application of convolutional neural networks in mammography serves multiple recognition purposes (Abdelhafiz et al., 2019). The first one is lesion classification, in which the model assigns a class to a ROI including a lesion. This can be two classes (benign/malignant), or three classes (bening/malignant/no tumor). Also, classification can lead to a probability of microcalcifications being present in the image. Lesion classification is the most used application in research up to this date (Wong et al., 2020). A second application is image classification, which takes the whole mammogram as input and includes the step of determining whether there is a lesion present in the image. This results in a class for the image of healthy or 'cancer-containing'. The third application is risk assessment, in which the risk to develop breast cancer in short-term is calculated, based on mammograms. High density of the breast tissue can generate an increased risk of breast cancer. Women with extremely dense breasts have a relative risk of 2.1 times higher than women with average breast density. The reason for this probably lies in the masking effect (breast tissue covers the tumorous tissue), too much glandular tissue and the nature of the breast tissue itself. The fourth application is lesion localization, where the model determines where exactly in the image an object, in this case the lesion, is placed. Lesion localisation is usually more complex when breast tissue is dense (Wong et al., 2020). Models with this purpose need a well-annotated dataset. Classification and localization is often combined in a model, such that a fixed number of lesions in an image is classified and located, this is called multi-class localization. Image retrieval and super resolution image reconstruction are the two last applications of convolution neural networks. Until this moment, CNNs are used in research, but they are not widely applied in clinical practices (Mello-thoms, 2020).

Year	Reference	Model	Database (number of images)	Application
2020	(Wu et al., 2020)	ResNet	Private (1.001.093)	Image classification
2020	(Mckinney et al., 2020)	RetinaNet, MobileNet, ResNet	Private UK (25.856) Private US (3.097)	Image classification
2019	(L. Shen et al., 2019)	Resnet, VGG-16	Private (1.001.093)	Image classification
2019	(Y. Shen et al., 2019)	Resnet	DDSM (2478)	Localization & classification
2019	(Singh et al., 2019)	cGAN CNN with 3 conv layers	INbreast(410) DDSM(1168) Private(300)	Tumor segmentation & shape classification
2019	(Savelli et al., 2019)	4 CNNs 4-10 conv layers	INbreast(410)	Detection of microcalcifications
2019	(R. Shen et al., 2019)	VGG-16	DDSM (2223)	Detection & segmentation
2019	(Duggento et al., 2019)	multiple CNNs 2-5 layers 78-612 input size 4-64 kernels 2-4 pooling size 1-3 fully connected layers 200-5 nodes in layer	DDSM(1696)	Image classification
2018	(Chougrad et al., 2018)	VGG-16 ResNet Inception	DDSM (>2600) BCDR (600) INbreast(410)	Image classification
2018	(Geras et al., 2017)	MV-DCN	Private (886.437)	Image classification
2017	(Al-masni et al., 2017)	YOLO	DDSM (600)	Detection & classification
2017	(Carneiro et al., 2017)	Alexnet	INbreast (410) DDSM (680)	Classification
2017	(Kooi et al., 2017)	VGG	Private (45.000)	Localization & classification
2016	(Hwang & Kim, 2016)	self-transfer-learning	DDSM (10363) MIAS (322)	Localization & classification
2016	(Dhungel et al., 2016)	LeNet & R-CNN	INbreast (410)	Detection, segmentation and classification
2016	(Sun et al., 2016)	CNN with 3 pairs conv layers	Private (3748)	Image classification

Table 3.1: Overview of research on the application of convolutional neural networks for mammography

3.2.2. Challenges in deep learning for mammography

The current challenges in the field of deep learning for mammogram evaluation are the following (Abdelhafiz et al., 2019):

- Localization of tumors is a complicated task. There is no model architecture yet that meets all requirements. For example, patch-based CNNs usually results in too many false-positives, r-CNNs are time-consuming and require much memory. The YOLO method requires less computation and memory, but results in lower accuracy.
- Limited data for training prevents a high accuracy of CNNs.
- There is usually an imbalance in positive and negative classes in the training data, since more images are healthy. This could be beneficial in training a model, to bias the model in the direction of the more common class. On the other hand, it can complicate an accuracte detection of negative classes.
- The size of lesions can be different among mammograms. Therefore, the ROI's are usually rescaled before a model is trained based on this data. Sometimes, this can affect the visibility of the lesion and the ability for the model to learn the characteristics.
- There are (almost) no annotated datasets available for mammograms. Usually there are only binary labels that tell the class, which can also be used as input for the CNN. However, information about the location of the abnormalities could improve the model.
- False positives still occur too often with CNNs. Research is done to improve this by combining prior

images from the same patient with the current images.
- Pre-processing filters can greatly improve the results of CNNs. Deciding on the best pre-processing methods is still under investigation.

3.3. Performance measures

To compare the performance of different classification models, multiple performance measures are taken into account, like described in chapter 3. This section describes how can be determined whether a model is significantly better at such a classification, and how the results are compared to the segmentation by the radiologist.

3.3.1. Evaluation metrics

To evaluate the performance of breast cancer detection models created with deep learning, different quantitative metrics can be used. The accuracy, sensitivity, specificity and precision can all be calculated from the confusion matrix. A confusion matrix shows the number of instances that is predicted (in)correctly by the model, and whether they belong to the class of cancer-containing or cancer-free images. A confusion matrix is shown in Table 3.2 and the calculation of the metrics are all shown in Table 3.3. The abbreviations in the evaluation metrics can be found in the confusion matrix.

		Predicted class	
		1 (cancer-containing)	0 (cancer-free)
Real class	1 (cancer-containing)	True Positive (TP)	False Negative (FN)
	0 (cancer-free)	False Positive (FP)	True Negative (TN)

Table 3.2: Confusion matrix

Metric	Formula
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$
Sensitivity	$\frac{TP}{TP+FN}$
Negative Predictive Value	$\frac{TN}{TN+FN}$
Specificity	$\frac{TN}{TN+FP}$
Precision	$\frac{TP}{TP+FP}$

Table 3.3: Evaluation metrics

Another way to interpret the results is to look at the Receiver Operating Characteristic (ROC), which plots the sensitivity against one minus the specificity for changing values of the threshold which is used to determine a class from a probability. For threshold values $z \in [0,1]$, the True Positive Rate (sensitivity) and the True Negative Rate (specificity) are defined as follows, where $X_1,,X_m$ are the predictions of the observations that belong to one class and $Y_1,,Y_n$ are the predictions of the observations that belong to the other class. :

$$TPR(z) = \frac{1}{m}\sum_{i=1}^{m} 1\{X_i \geq z\} \tag{3.7}$$

$$TNR(z) = \frac{1}{n}\sum_{j=1}^{m} 1\{Y_j < z\} \tag{3.8}$$

The AUC is the area under this ROC curve. It can be calculated as follows:

$$\hat{\theta} = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} \psi(X_i, Y_i), \text{ where } \psi(X_i, Y_i) = \begin{cases} 1 & Y < X, \\ 1/2 & Y = X \\ 0 & Y < X \end{cases} \tag{3.9}$$

An example of an ROC curve is shown in Figure 3.2.The Area Under the Curve (AUC) is calculated from this graph and represents the area under the ROC curve, ranging between 0 and 1. In the figure this is the pink area. An AUC score of 1 indicates that the model can perfectly separate negative and positive examples, an AUC score of 0.5 equals the score of random classification.

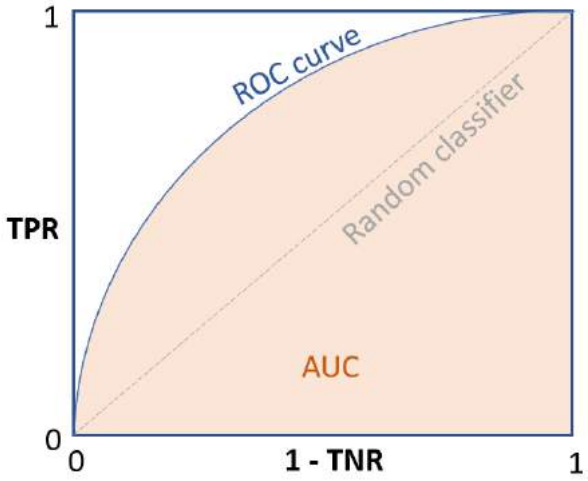


Figure 3.2: Receiver operating characteristic curve

3.3.2. DeLong test

To demonstrate that the results from one classification model on a test set is significantly better than the performance of another model on the same test set, DeLong's test can be used (DeLong et al., 1988). From the test a p-value can be obtained for whether the two AUCs from the models are significantly different. The AUC is defined as previously stated. Now when there are K classifiers,we can define

$$\hat{\theta}^k = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\psi(X_i^k, Y_i^k) \tag{3.10}$$

Now set $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1 \dots \hat{\theta}_k)^\intercal \in \mathbb{R}^K$ as the empirical AUCs and $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1 \dots \hat{\theta}_k)^\intercal \in \mathbb{R}^K$ as the true AUCs. Set $\mathbf{L} \in \mathbb{R}^K$ a vector of coefficients. In the case of two classifiers, $K = 2$ and $L = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$. Also, define two $K \times K$ matrices with the (r,s)th element as follows:

$$(S_{10})_{rs} = \frac{1}{m-1}\sum_{i=1}^{m}\left[V_{10}^r(X_i) - \hat{\theta}^r\right]\left[V_{10}^s(X_i) - \hat{\theta}^s\right] \tag{3.11}$$

$$(S_{01})_{rs} = \frac{1}{n-1}\sum_{j=1}^{n}\left[V_{01}^r(X_i) - \hat{\theta}^r\right]\left[V_{01}^s(X_i) - \hat{\theta}^s\right] \quad \text{, with} \tag{3.12}$$

$$V_{10}^r(X_i) = \frac{1}{n}\sum_{j=1}^{n}\psi(X_i^r, Y_j^r) \tag{3.13}$$

$$V_{01}^r(X_i) = \frac{1}{m}\sum_{i=1}^{m}\psi(X_i^r, Y_j^r) \tag{3.14}$$

Then,

$$\frac{\mathbf{L}^T\hat{\boldsymbol{\theta}} - \mathbf{L}^T\boldsymbol{\theta}}{\sqrt{\mathbf{L}^T(\frac{1}{m}S_{10} + \frac{1}{n}S_{01})\mathbf{L}}} \tag{3.15}$$

has a standard normal distribution. Then, to compare two AUCs, the null hypothesis $H_0 : \theta_1 = \theta_2$ is tested. If This hypothesis is true, the distribution of

$$\frac{\mathbf{L}^T\hat{\boldsymbol{\theta}} - \mathbf{L}^T\boldsymbol{\theta}}{\sqrt{\mathbf{L}^T(\frac{1}{m}S_{10} + \frac{1}{n}S_{01})\mathbf{L}}} = \frac{\hat{\theta}_1 - \hat{\theta}_2}{\sqrt{\mathbf{L}^T(\frac{1}{m}S_{10} + \frac{1}{n}S_{01})\mathbf{L}}} \tag{3.16}$$

is the standard normal distribution. If this result in absolute value is too high, the null hypothesis can be rejected and it can be believed that there is a difference between the two AUCs.

3.3.3. Comparison of deep learning to radiologist reviews

In a paper by Wong et al. (2020), 33 studies that use CNNs in mammography are compared with four studies with radiologists on some of these metrics, the results are shown in Table 3.4. From this table it can be observed that radiologists generally have a higher specificity than CNNs. For the sensitivity, CNNs and radiologists have comparable results. It must be noted that the datasets used are particularly different. The

	Mean		Median		Min		Max	
	CNN	radiologist	CNN	radiologist	CNN	radiologist	CNN	radiologist
AUC	0.851	-	0.876	-	0.540	-	0.996	-
Specificity	0.851	0.905	0.890	0.908	0.660	0.889	0.970	0.988
Sensitivity	0.833	0.795	0.899	0.916	0.440	0.600	0.971	0.899
Total accuracy	0.883	-	0.930	-	0.660	-	0.990	

Table 3.4: Comparison between studies that use CNNs and studies with radiologists (rad) (Wong et al., 2020)

studies using CNNs often use a small publicly available dataset like DDSM or MIAS, while the studies concerning radiologists use a more clinical setting with images from a screening program. Moreover, specific conditions may be applied for selection of the images. The lowest sensitivity score for radiologists of 0.600 is the results of mammography screening to a population of women with symptoms, that do not match the normal symptoms for breast cancer. The comparison of deep learning models with each other and the comparison to radiologist requires standardized datasets.

4

Data description

Two datasets consisting of mammographic images are used in this study, which are managed by two medical institutes: The Erasmus Medical Centre and the Karolinska Institute. For clarification, one mammography examination consists of multiple mammography images.

4.1. Data from the Erasmus MC

Patients who are treated in the Erasmus MC, can be redirected from the population screening program, or have come to the hospital on their own because of other reasons like palpable abnormalities. Furthermore, they can be redirected from other hospitals for treatment at the Erasmus MC. In all these situations, image from their mammography examinations are available in the Erasmus MC database.

4.1.1. Description

The dataset from the Erasmus MC consists of mammography examinations from 197 patients diagnosed with ILC. Of those patients, 45 have the Erasmus MC listed as their first hospital. Some of these patients have multiple mammography examinations available in the database on different dates. In this situation, the mammography examinations is taken, that is made before the diagnosis with the shortest time in between study and diagnosis. One of these 45 patients only had a mammography examination available from after diagnosis and after surgery, so this examination is taken out of the dataset. Of these 44 examinations, 9 examinations contain two images from one breast, and the other 35 examinations contain four images, two from each breast.

For the patients who were diagnosed and treated in another hospital first, the dataset is less structured. Mammography examinations with images from 152 patients are available, but they do not always include the correct views. From these 152 examinations, 34 contain only images that are not usable. This leaves 118 examinations. From these 118 examinations, 61 contain four images from the four views. 43 contain two images from one breast, three examinations contain three images and one examination contains just one image. The total dataset therefore contains 530 images from 162 examinations. Of these images 180 are used in the experiments, of which 112 contain a tumor. These are the images that are reviewed by the breast radiologist, like described in subsection 4.1.2.

The examination date of all mammograms is after 2006. This selection criterion is used, to make sure the mammograms are digital. In figure Figure 4.1 a histogram is displayed, showing when the mammogram examinations in the final selection in this dataset were made.

4.1.2. Labels and segmentation

To determine the laterality (in which breast is the tumor located) and the location of the tumor within the breast, a breast radiologist from the Erasmus MC reviewed a subset of the images in the dataset. This is done for all 44 examinations from patients from the EMC, and for 11 patients from other hospitals, which results in a total of 55 segmented tumors. To perform this segmentation task, the radiologist used multiple resources. The radiology report, together with reports from pathology and available MRI scans were used to determine

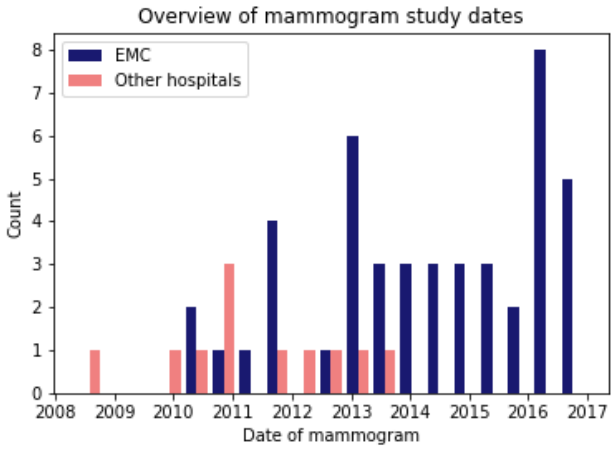


Figure 4.1: Overview of dates of mammograms in EMC dataset

(a) Mammogram (b) Drawing radiologist (c) Final segmentation

Figure 4.2: Example of mammogram from EMC dataset, together with its segmentation and processed segmentation

the true laterality and mark the location of the tumor on the mammographic image. In this way, more information was obtained than would have been possible with the mammographic images alone.

For the segmentation, some operations were performed to transform the radiologist's drawing to a segmentation which include all tumor pixels. First, holes in the lines are filled, followed by filling in the insides of the contour. An example of this is shown in Figure 4.2. The algorithm to do this is given in section A.2.

4.2. Cohort of Screen-Aged Women

The other dataset that was used in this study is the Cohort of Screen-Aged Women (Dembrower et al., 2020). This dataset, developed and managed by the Karolinska Institute in Stockholm, consists of mammographic images from the population screening of women in the Stockholm region between 2008 and 2015. The women were between 40 and 74 years old, at the time of screening. The dataset includes images and - for a selection of images - tumor masks on pixel-level by a radiologist. Furthermore, metadata is given for all patients, including the breast cancer diagnosis, histology, tumor size, lymph node status and receptor status. This information was obtained from pathological examination. Furthermore, the review of the radiologists is given.

4.2.1. Labels

For each women, one or multiple mammography examinations are available. In the case of multiple examinations, the time between the examinations is 18 to 24 months each time. For women where a tumor developed during the screening period, there are no more mammograms after this development. The latest

available mammogram is therefore either the mammogram that is used for the tumor diagnosis, or the mammogram that the radiologists reviewed as healthy, but a tumor developed soon after this screening. Therefore, there is a high possibility of the tumor already being present on the mammogram. For the examinations, prior to this latest examination, it cannot be determined certainly whether there was already a tumor present or not.

The labels of the images are determined as follows. The examinations that belong to women that did not develop a tumor during the screening period, are labeled as healthy. For the women that did develop a tumor during the screening period, both images, belonging to the breast where the tumor developed from the latest available mammography examination are labeled as tumor images. The two images from the other breast - where no tumor developed - are labeled as healthy. The earlier examinations of these women are left out of this study, because of the uncertainty of the tumor status in the images.

In Table 4.1 an overview is given of the distribution of examinations with breast cancer and healthy images. To clarify, the examinations during diagnosis consist of two images from the breast with the tumor and two images of the healthy breast from the same woman. This results in 180 images that contain an ILC tumor, during diagnosis. Also, a distinction is made between different histology types, since the focus of this study is on ILC. The images that are used include the mammograms from women with no tumor development, and the images during diagnosis from women with tumor development.

Number of mammograms			
Women with no tumor development	**Women with tumor development**		
		During tumor development	Before tumor development
22803	ILC	95	105
	IDC	605	673
	Other	172	175

Table 4.1: Number of mammography examinations for patients with and without breast cancer, divided into breast cancer histology

5

Methods and experiments

This chapter describes the methods that are used in the study. Furthermore, it describes the steps that are made in the experiments and the data that is used. The first section explains the performance measures that are used, because they will come back later

The first section describes image processing methods, the second describes deep learning methods, then the methods to predict breast density, and lastly the used performance measures are explained.

5.1. Exploratory research using conventional image processing techniques

Image processing techniques are applied to mammography images to make features more visible, detect the tumor edges, segment objects and detect masses. This section describes which methods are used in this study, and how the effects of these methods are quantified. The data used in these analyses is the EMC set.

5.1.1. Methods used

The following seven methods are selected to test on the dataset, which are explained in detail in chapter 2:

- M_1: Adaptive histogram equalization
- M_2: Histogram equalization
- M_3: Threshold segmentation
- M_4: Local threshold segmentation
- M_5: Sobel transform
- M_6: Unsharp masking
- M_7: Wavelet transform

These methods are selected, because they have successfully been applied to mammography in previous studies. Moreover, they all create an output image from an original image, which enables a comparison between the methods. The image after application of the image processing method M_m is as follows:

$$I_m = M_m(I) \quad \text{with } m \in \{1,2,3,4,5,6,7\} \tag{5.1}$$

5.1.2. Selecting healthy and tumor tissue

To determine the effects of the methods on the mammography images, and specifically the tumor area, the tumor area and a corresponding healthy area have been determined. The healthy area is determined, because this enables a comparison between the effect on the tumor area and the effect on a healthy area. The tumor area follows from the segmentation, made by the breast radiologist and the operations to transform this into the correct mask, like described before. This tumor area is called R_{tumor}, with $(x,y) \in R_{tumor}$ if $T(x,y) = 1$ with T the binary tumor mask corresponding to the same mammography image. The corresponding healthy area, which will be called $R_{healthy}$, is determined in two ways. The first way is used, when there are also images available from the healthy breast from the same mammography study. In this case, the tumor mask is flipped horizontally and added to the healthy corresponding image. Therefore the elements of $R_{healthy}$ are $(-x,y)$ if $T(x,y) = 1$. This means that for a *L-CC* image with a tumor, the healthy area is the flipped tumor mask from this image, applied to the *R-CC* image. The same holds for the *L-MLO* image, which has a corresponding

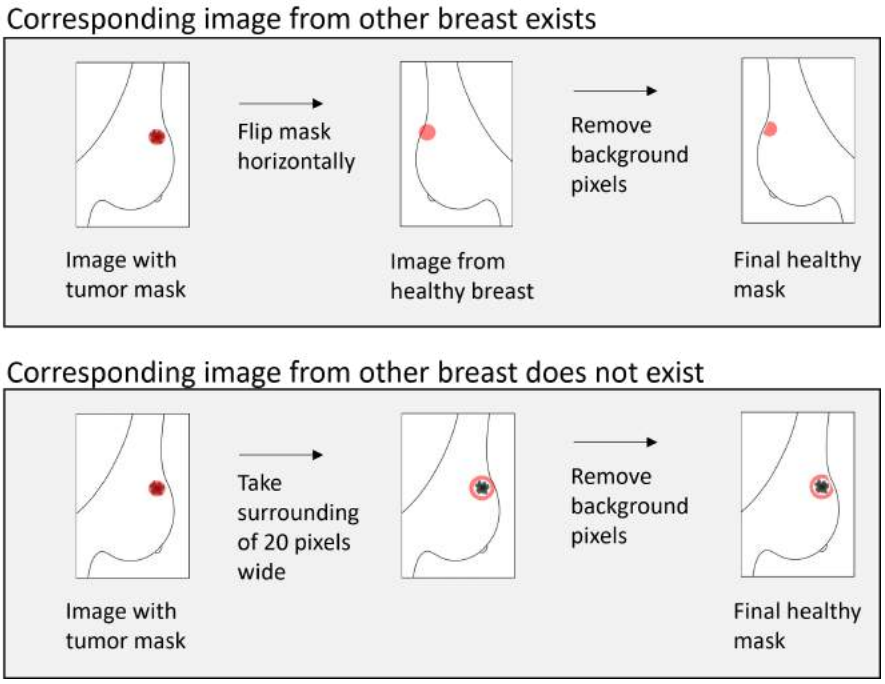


Figure 5.1: Illustration of production of healthy mask

healthy area from the *R-MLO* image. This corresponding image is called I_c. Sometimes, this mask could fall out of the breast tissue area, and cover some background pixels. In this case, the pixels outside of the breast are removed from $R_{healthy}$. In some cases, there are just images available from the breast with the tumor present, and not from the corresponding healthy breast. In these situations, the healthy area is taken from a region in the same image, outside the tumor area. A region of 20 pixels outside the tumor mask is selected as healthy area. In the same way as before, when this area covered pixels outside the breast area, they were removed from the area. An illustration of this is shown in Figure 5.1.

5.1.3. Analyses

For each processing method m, the pixel values in the tumor and healthy area after image processing are determined, with (x, y) being coordinates in the image. For the $healthy-pixel-values$, it depends which image is used to determine the healthy area. When the healthy area is taken from the corresponding healthy image, Equation 5.3 applies. When the healthy area is taken from the region outside the tumor in the same image, Equation 5.4 applies.

$$tumor-pixel-values_m = \{I_m(x,y) \mid (x,y) \in R_{tumor}\} \tag{5.2}$$

$$healthy-pixel-values_m = \{I_{c_m}(x,y) \mid (x,y) \in R_{healthy}\} \tag{5.3}$$

$$healthy-pixel-values_m = \{I_m(x,y) \mid (x,y) \in R_{healthy}\} \quad \text{for } m \in \{1,2,3,4,5,6,7\} \tag{5.4}$$

Then, $tumor-pixel-values_m$ get a true label 1, and $healthy-pixel-values_m$ get a true label 0. The next step is to determine how much these pixel values differ from each other. To do this, the AUC value is used, which is defined in subsection 3.3.1. This is done for all images in the dataset, and all resulting AUC values have been compared.

5.2. Deep learning

The deep learning experiments performed in this research are focused at determining the performance of existing trained models on ILC and general breast cancer datasets. Furthermore, experiments are performed to improve the performance of one of these models. First, the models are described, and thereafter the experiments are explained. Lastly, metrics to assess the performance are mentioned.

5.2.1. Pre-trained models

There are six models using convolutional neural networks that are used for the first evaluation. The architecture and data used for training for each model are briefly described in Appendix B. Figures to clarify the architectures, from the original papers, are also added there. The following section focuses on a more elaborate explanation of the model that is used in the experiments that follow later, which is named GMIC (Y. Shen et al., 2019). This model is chosen for transfer learning, due to its structure which enables training with labels on image-level and high performance in the first experiments.

To determine the models in their ability to classify mammograms into healthy and tumor images, they are all tested on the two datasets described in chapter 4. For the EMC dataset, this means all images are taken into account. For the CSAW dataset, all images containing a tumor are used, combined with 2000 healthy mammography studies. The results of these experiments are given in subsection 6.2.1. For the models that have an outcome prediction for malignant as well as benign abnormalities, only the malignant prediction is taken into account for this evaluation.

5.2.2. Description GMIC model

Figure 5.2 shows the outline of the model, created by Y. Shen et al., 2019. The structure consists of three Modules: the Global, Local and Fusion Module. The goal of the model is to predict two labels for one image $\mathbf{x}$, one for the probability of a benign finding in a mammogram, and one for the probability of a malignant finding in a mammogram. So, the label can be written as

$$\mathbf{y} = \begin{bmatrix} y^b \\ y^m \end{bmatrix} \text{ with } y^b, y^m \in \{0,1\}. \tag{5.5}$$

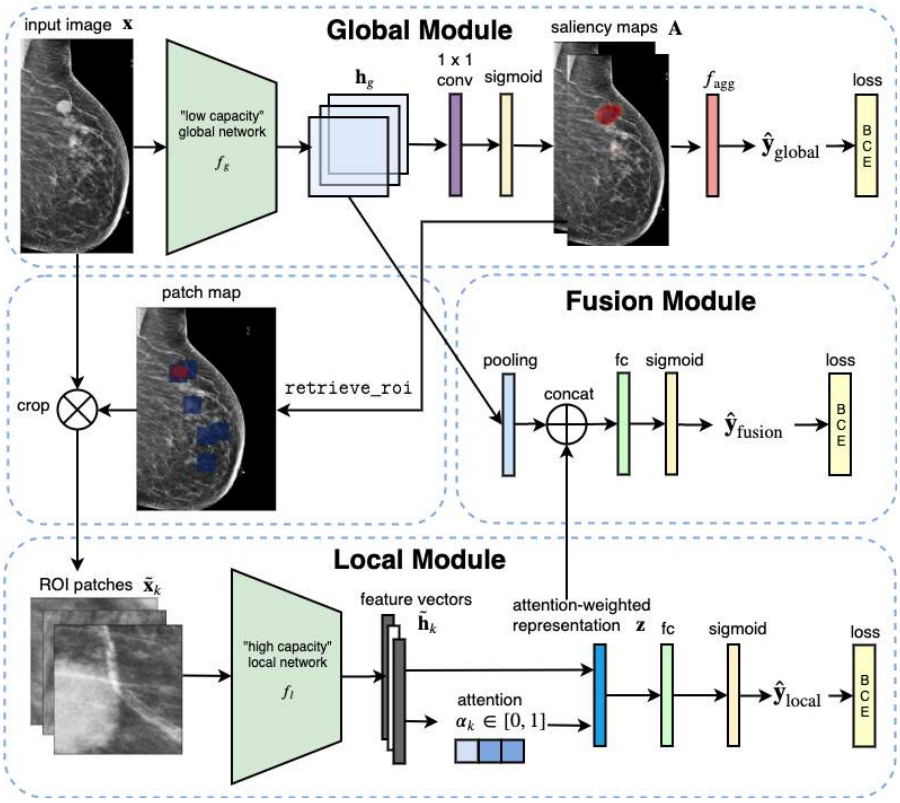


Figure 5.2: The architecture of the GMIC model (Y. Shen et al., 2019)

An input image is first fed to the Global Module. The Global Module operates on a down-sampled version of the mammogram and uses a network for feature extraction f_g, to create a feature map from the image:

$\mathbf{h}_g = f_g(\mathbf{x})$. Two saliency maps $\mathbf{A}^b, \mathbf{A}^m$ are created, through a 1x1 convolution layer with sigmoid non-linearity: $\mathbf{A} = \text{sigm}(\text{conv}_{1x1}(\mathbf{h}_g))$. From these saliency maps, regions of interest are selected to feed to the Local Module. The saliency maps are created using image level labels, without any information on the location of the tumor in the image. Therefore, an aggregation function is defined, to convert the saliency maps to image level labels. This aggregation function balances between global average pooling and global max pooling, by taking the average of the maximum of a percentage of pixels in the saliency map. With H^+ as the locations for the maximum values in the saliency maps, the aggregation function can be defined as follows:

$$\hat{\mathbf{y}}^c_{global} = f_{agg}(\mathbf{A}^c) \tag{5.6}$$

$$f_{agg}\left(\mathbf{A}^c\right) = \frac{1}{|H^+|} \sum_{i,j \in H^+} \mathbf{A}^c_{i,j} \tag{5.7}$$

Using the saliency maps, the regions of interest are selected to feed into the Local Module. The algorithm for this is given in section A.1. These patches, $\tilde{\mathbf{x}}_k$ are then fed into the feature extraction network of the Local Module f_l. This creates feature vectors $\tilde{\mathbf{h}}_k = f_l(\tilde{\mathbf{x}}_k)$. These vectors are added together by a function f_a, which means $\mathbf{z} = f_a(\{\tilde{\mathbf{h}}_k\})$. The functions f_a is defined as

$$\mathbf{z} = \sum_{k=1}^{K} \alpha_k \tilde{\mathbf{h}}_k \tag{5.8}$$

with

$$\alpha_k = \frac{\exp\{\mathbf{w}^\top(\tanh\left(\mathbf{V}\tilde{\mathbf{h}}_k^\top\right) \odot \text{sigm}\left(\mathbf{U}\tilde{\mathbf{h}}_k^\top\right))\}}{\sum_{j=1}^{K} \exp\{\mathbf{w}^\top(\tanh\left(\mathbf{V}\tilde{\mathbf{h}}_j^\top\right) \odot \text{sigm}\left(\mathbf{U}\tilde{\mathbf{h}}_j^\top\right))\}} \tag{5.9}$$

The final layer of the Local Module is fully connected and produces a predictions score from the feature vector, using sigmoid activation.

$$\hat{\mathbf{y}}_{local} = \text{sigm}(\mathbf{w}^\top_{local}\mathbf{z}) \tag{5.10}$$

Lastly, the information that is retreived in both the Global and Local Modules is added together in the Fusion Module, to make a final prediction for the image. This is done by appying global max pooling to $\mathbf{h}_g$ and concatenating this to $\mathbf{z}$, after which a fully connected layer is placed with sigmoid activation:

$$\hat{\mathbf{y}}_{fusion} = \text{sigm}(\mathbf{w}_f[GMP(\mathbf{h}_g), \mathbf{z}]^\top) \tag{5.11}$$

The loss function that is used during training of the GMIC model consists of the three predictions outputs y_{local}, y_{global} and y_{fusion}, and a regularizing term on $\mathbf{A}^c$ to induce sparsity on the saliency maps. For $\mathbf{y}$ being the true output value and $\hat{\mathbf{y}}$ the predictions from the model, the loss function is as follows, where BCE stands for Binary Cross Entropy.

$$L(\mathbf{y}, \hat{\mathbf{y}}) = \sum_{c \in \{b,m\}} \text{BCE}(\mathbf{y}^c, \hat{\mathbf{y}}^c_{local}) + \text{BCE}(\mathbf{y}^c, \hat{\mathbf{y}}^c_{global}) + \text{BCE}(\mathbf{y}^c, \hat{\mathbf{y}}^c_{fusion}) + \beta \sum_{(i,j)} |\mathbf{A}^c_{i,j}| \tag{5.12}$$

The big advantage of the combination of the local and the global module, is that it allows the model to combine global features on a lower resolution, with local features from the full resolution. Usually a balance has to be found, between a high resolution and a smaller model, or a small resolution and a deeper or wider model. Moreover, by selecting K patches to run the Local Module, it is prevented to run this module on all patches from the mammogram, but only use the patches that are classified as high risk by the Global Module.

There are some preprocessing steps that are taken before using the data for training or for inference. These steps include cropping the mammogram by discarding the background, then resizing the image to 2944 × 1920 pixels, and finally normalizing the image to have a mean of 0 and a standars deviation of 1. To prevent overfitting, data augmentation is applied in the form of random cropping and size noise.

The dataset that is used to train this model is the NYU breast cancer dataset (Wu et al., 2019). This dataset contains 229,426 exams from 141,472 patients, with a total of 1,001,093 images (458,852 breasts). From these images, 985 breasts contain malignant findings and 5556 breasts contain benign findings. The exams are divided into subsets for training, validation and testing, of which the amounts are shown in Table 5.1. Images from one examination are put in the same set.

	Total exams	Exams with malignant findings	Exams with benign findings
Training	186816	857	4590
Validation	28462	66	610
Testing	14148	62	356

Table 5.1: Number of exams in the NYU breast cancer dataset (Wu et al., 2019)

Optimizer
The optimizer used in this model is the Adam optimizer (Kingma & Ba, 2015). This is an algorithm for optimization for gradient descent, that lately has been used broadly for deep learning applications in computer vision and natural language processing. The optimization algorithm in pseudo code is shown in Algorithm 1.

Algorithm 1 Adam optimization

Require: α: Stepsize
Require: $\beta_1, \beta_2 \in [0,1)$: Exponential decay rates for the moment estimates
Require: $f(\theta)$: Stochastic objective function with parameters θ
Require: θ_0: Initial parameter vector
$m_0 \leftarrow 0$ (Initialize 1st moment vector)
$v_0 \leftarrow 0$ (Initialize 2nd moment vector)
$t \leftarrow 0$ (Initialize timestep)
while θ_t not converged **do**
  $t \leftarrow t+1$
  $g_t \leftarrow \nabla_\theta f_t(\theta_{t-1})$ (Get gradients w.r.t. stochastic objective at timestep t)
  $m_t \leftarrow \beta_1 \cdot m_{t1} + (1\beta_1) \cdot g_t$ (Update biased first moment estimate)
  $v_t \leftarrow \beta_2 \cdot v_{t1} + (1\beta_2) \cdot g_t^2$ (Update biased second raw moment estimate)
  $\hat{m}_t \leftarrow m_t / (1-\beta_1^t)$ (Compute bias-corrected first moment estimate)
  $\hat{v}_t \leftarrow v_t / (1-\beta_2^t)$ (Compute bias-corrected second raw moment estimate)
  $\theta_t \leftarrow \theta_{t-1} - \alpha \cdot \hat{m}_t / (\sqrt{\hat{v}_t} + \epsilon)$ (Update parameters)
end while
return θ_t (Resulting parameters)

5.2.3. Transfer learning

Weiss et al. (2016) define transfer learning as follows: A target learning task $\mathbb{T}^T$ over a target domain $\mathbb{D}^T$, using already acquired knowledge of another source learning task $\mathbb{T}^S$ over a source domain $\mathbb{D}^S$. Moreover, usually $\mathbb{T}^T \neq \mathbb{T}^S$ and $\mathbb{D}^T \neq \mathbb{D}^S$. In the experiments in this study, $\mathbb{D}^S$ are all mammographic images and $\mathbb{T}^S$ is the task of classifying into cancer-free and cancer-containing. $\mathbb{D}^T$ are a combination of healthy images and images containing ILC, and $\mathbb{T}^T$ is the task of classifying into cancer-free and ILC-containing images. This means, there is a model trained on a mammography dataset containing healthy images and tumor images from all tumor types, that is finetuned using an ILC mammography dataset containing healthy images and tumor images with ILC, with the goal of improving the detection of ILC. This is visualized in Figure 5.3, where the darker parts are the parts that are new in comparison to the original model. The model that is used for transfer learning is GMIC, because it has the highest performance of the end-to-end trainable models, in the first experiments.

There are two different approaches of transfer learning that are used in the experiments:

1. **Feature extraction**
   In this approach, the source model is used as feature extractor. This means all layers before the final layers are frozen. These final layers are the fully connected layers that in the end decide on the classification of the image. All layers before this one, ultimately have the goal of creating well defined feature vectors for the input images. The motivation for this approach lies in the similar features of ILC and other malignant tumor types, which could indicate that the same features could be used. Moreover, this approach requires less training time, since not all weights need to be updated. Furthermore, the chance of overfitting is small, because the model cannot incorporate features that are specific to the training set.

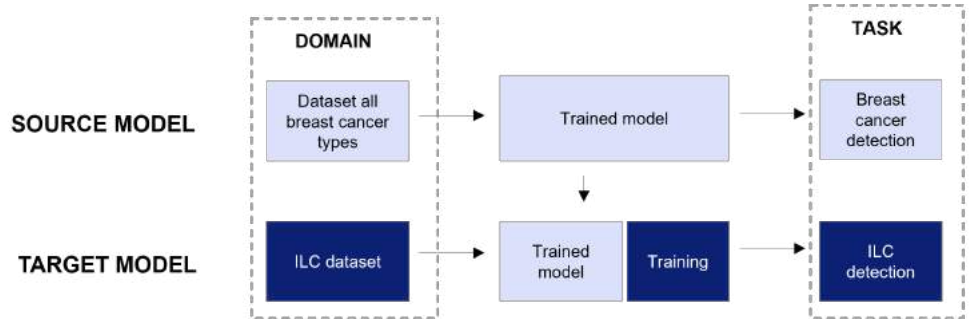


Figure 5.3: Transfer learning

2. **Weight initialization**
   The weight initialization approach does not make a difference between weights in layers of the model. In this experiment, all weights are updated during transfer learning, including the layers belonging to the convolutional layers that create the feature vectors. This could yield a higher accuracy, when the features to classify ILC are different than the features to classify other tumor types.

During the transfer learning experiments, 10 models are trained with different learning rates between 10^{-6} and 10^{-5}. Training takes place for 40 epochs. In the original training of the model, the learning rate was varied between $10^{-5.5}$ and $10^{-4.5}$ and training went on for 50 epochs. The decision to use a lower learning rate and less epochs stems from the transfer learning approach, where a big parts of the weights in the model are already set and transfer learning should not have too much influence on the weights, to not disturb the created structure.

The loss function that is used is the same loss function as in the original paper, like described in subsection 5.2.1. In each of the 40 epochs the loss is calculated over the training set and the test set. After 40 epochs the final model is saved, together with the model during the epoch where the loss over the test set was the lowest. This is done to prevent overfitting of the model. The results of these experiments are shown in subsection 6.2.2.

5.2.4. Training and test data

The training set should contain images with ILC and healthy images, like described in the previous section. The data used for training are all part of the CSAW dataset, see chapter 4. The mammograms that are used firstly include the latest mammograms from ILC patients. Following Table 4.1 this resulted in 95 mammography sessions with four images each. Two images belong to the breast where ILC is located, and two images belong to the other healthy breast. Therefore, this results in 190 images with ILC label and 190 images with healthy label. These 190 ILC images are split into a set for training and a set for testing, following a 80%/20% distribution. This results in 152 training images and 38 testing images. The healthy corresponding images from the same examination, but showing the other breast are added to the same set. Secondly, mammograms are added from women who did not develop breast cancer as healthy images. There are 22803 mammography examinations from women that did not develop breast cancer. Of these examinations, a random subset of 4,000 is taken for training, resulting in 16,000 healthy images. Another random subset of 1,000 examinations is taken and the 4,000 images from these examinations are added to the test set. In this way, images from the same evaluation are in the same set. Also, it has been ensured that examinations are not both in the training and test set. To be able to test the results also on a dataset with all tumor types, 20% of the images containing other tumor types are randomly selected and added to the test set, creating a new test set called *CSAW test set - all*. Since there are 777 examinations with other tumor types, this results in 155 examinations containing 310 images with a tumor and 310 images without a tumor. An overview of the different sets is shown in Table 5.2. For the sake of completeness, the EMC dataset is also added here.

During every epoch - which is an iteration of the training process - all 152 ILC training images are shown to the model, and a random selection of the same amount of healthy images. The model is trained for 40 epochs. Besides labels for malignant findings, the GMIC model requires labels for benign findings in mammograms, since the prediction for this is also built into the structure of the model. When trying to predict

Subset name	Number of healthy images	Number of images containing a tumor	
		ILC	IDC and other
CSAW training set	16,152	152	0
CSAW test set - all	4,348	38	310
CSAW test set - ILC	4,038	38	0
EMC set	68	112	0

Table 5.2: Overview of training and test sets

ILC, the focus should be on malignant findings. Furthermore, there are no labels given in the CSAW data as well as in the EMC data for benign findings. Therefore, the label for benign findings is set to 0 in all images used for training.

5.2.5. Training from scratch

To determine the added value of transfer learning in the experiments, additional models are trained where the weights from the original model are not used as initialization. Instead, the weights are initialized with the default method from pytorch. In this method, the distribution from which a weight is randomly sampled is dependant on the type of layer of that weight. For a 2D convolutional layer with an input size of n, the weights are sampled from a uniform distribution between $-\frac{1}{\sqrt{n}}$ and $\frac{1}{\sqrt{n}}$.

The experiments where the model is trained from scratch use the same data from the CSAW dataset for training as the transfer learning experiments. The same values are used for the learning rate and the number of iterations. Moreover, the evaluation is done in the same way, using the testing set from the CSAW and the EMC datasets. Initialization of the model with the weights from the original study does not add much information, when the performance of these models that are trained from scratch is comparable to the transfer learned models.

5.2.6. Pre processing

An addition to the transfer learning approach is done, testing multiple image processing methods as preprocessing during training and testing of the models. The methods that are selected for this experiment, are derived from the results in section 6.1. Following these results, histogram equalization and unsharp masking are found as potential methods that could have a beneficial effect on ILC regions, making them clearer. The method of transfer learning that is chosen here is *weight initialization*, because the preprocessing methods change the structure of the image, and therefore the feature extraction should also be adjusted. Moreover, this decision is based on the more promising results of *weight initialization* in comparison with *feature extraction*, like described in subsection 6.2.2.

5.3. Breast density prediction

To be able to relate the results of the tumor prediction to the density of the breast tissue, an estimation of this density should be made. This is done using a prediction model of Wu et al. (2018). This paper describes two classifiers. The first one uses features based on the histogram of the pixel values in an image. The second one is a convolutional neural network, which automatically extracts features from the image. This convolutional neural network is used in this study, since it showed the best results in comparing the predicted labels to the labels set up by radiologists in the study by Wu et al. (2018). The model uses the four views of a mammogram as input, and produces a four-way classification for the four density categories, described in section 1.4. The exam is assigned the category belonging to the highest prediction value.

This information is used in different ways of evaluation of the results:

1. The images are split into non-dense (density category 0 and 1) and dense (density category 2 and 3) images, and the results of the model are evaluated on these two groups separately, to use breast density as a factor which can be used to assess how accurate a prediction is

2. Mammograms belonging to the most dense category are left out of the results. For the most dense group, there have been studies about the possibility of additional MRI screening, next to the regular

screening with mammography. Therefore, this evaluation shows what would be the result of this for the population screening.

5.3.1. Spatial analysis

In the model, the saliency map $\mathbf{A}_m$, resulting from the global module, can be interpreted as a heatmap, from which areas of concerns are determined. These heatmaps also offer additional insights in the parts of the image that are responsible for the models final prediction. To analyse this, these heatmaps are compared to the segmentation from the breast radiologist. There are two metrics used for this comparison. The first metric is Dice's coefficient, which reflects the similarity of a ground truth segmentation X and a predicted segmentation Y. First, the saliency map is transformed into a binary segmentation by setting a threshold value t:

$$Y(x,y) = \begin{cases} 1 & \text{if } \mathbf{A}_m(x,y) \geq t \\ 0 & \text{if } \mathbf{A}_m(x,y) < t \end{cases} \tag{5.13}$$

Because the optimal value of this threshold is yet unknown, this value is varied between 0 and 1, with steps of 0.05. Then, the coefficient is calculated as follows:

$$DSC = \frac{2|X \cap Y|}{|X| + |Y|} \tag{5.14}$$

Then, for each image, the optimal threshold will be determined by the value with the highest resulting DSC. Dice's coefficient not only gives information on the location of the segmentation, but also on the size, because it looks at the overlap between the ground truth and the segmentation. To leave this out and just consider the model's ability to locate the tumor correctly, the location of the maximum value of the saliency map is taken and compared to the true segmentation. When the maximum value of the saliency map falls within the true segmentation, the saliency map is considered to be correct. Otherwise, the saliency map is considered incorrect.

5.4. Comparison to radiologists

To find out the added value of the models compared to the opinion of the radiologists, the predictions made by the model are compared to the opinion of the radiologists. For this, the CSAW dataset is taken, because the radiologist review is given for all images. The features in the dataset referring to this review are given in Table 5.3.

Feature	Options
rad_r1 (review radiologist 1)	0: healthy 1: discuss
rad_r2 (review radiologist 2)	0: healthy 1: discuss
rad_recall (final review)	0: normal 1: recall

Table 5.3: Variables in CSAW dataset concerning the radiologist review

To determine the added value of the model prediction, *rad_recall* is used, because this variable reflects the final decision of the radiologists, and thus it includes most information from the radiologist reviews.

The comparison to radiologists is done, both for the *CSAW test set - ILC* and for the *CSAW test set - all*. The *CSAW test set - ILC* is used, because the goal of this study is explicitly to evaluate the method on ILC data. The reason to add the evaluation on the *CSAW test set - all*, is because it includes more tumor images and therefore the results may be more generally applicable.

6

Results

First, the results from the image processing methods are presented. Secondly, the results from the deep learning methods are presented and the relationship between breast density and correctness of the prediction are determined. Lastly, the results are compared to the reports of the radiologists.

6.1. Image processing

The selected image processing methods, like described in section 5.1 are performed on all images from the EMC dataset. The next section shows some examples of these images, and afterwards the comparison between the effect on the healthy and tumor tissue is presented.

6.1.1. Processed images

The resulting processed images are shown in Figure 6.1 for two example images. The image in the upper row shows a clearly visible tumor already in the original mammogram. In the image in the lower row, the tumor is not so well visible in the original image.

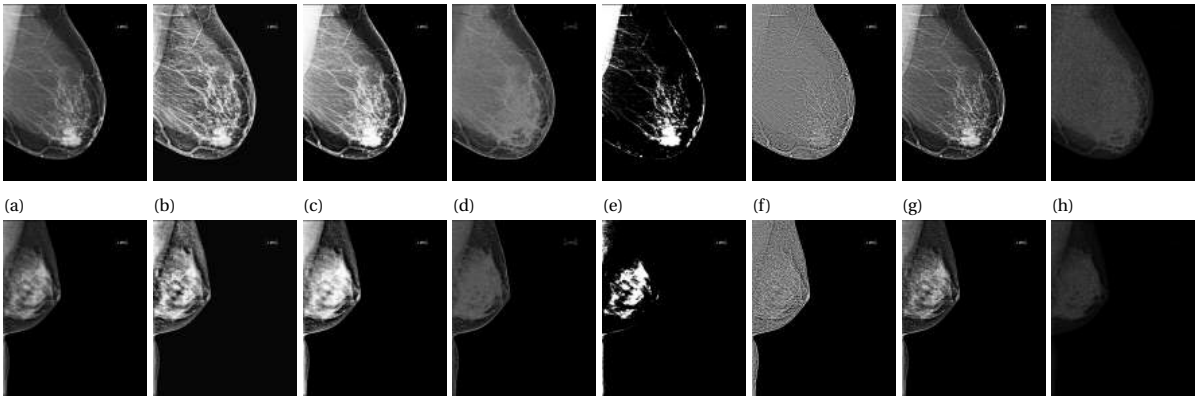


Figure 6.1: Resulting images for image processing methods for image where the tumor is clearly distinguishable (upper row) and not clearly distinguishable (lower low) : (a,i) original image (b,j) adaptive histogram equalization (c,k) histogram equalization (d,l) sobel transform (e,m) threshold segmentation (f,n) local treshold segmentation (g,o) unsharp masking (h,p) wavelet transform

For the image in the upper row it seems that adaptive histogram equalization, histogram equalization, threshold segmentation en unsharp masking preserve the contrast of the tumor against the background, while the sobel transform, local threshold segmentation and wavelet transform reduce this contrast. In the lower image adaptive histogram equalization, histogram equalization and unsharp masking seem to preserve the contrast between the tumor and the background tissue the most, while this does not hold for the other methods.

6.1.2. Comparison using AUC

To further investigate the effects of the image processing methods, the histograms of pixel values in the tumor area are compared to the histograms of a healthy counterpart, like described in subsection 5.1.2. Figure 6.2 shows this healthy and tumor area for the first example image.

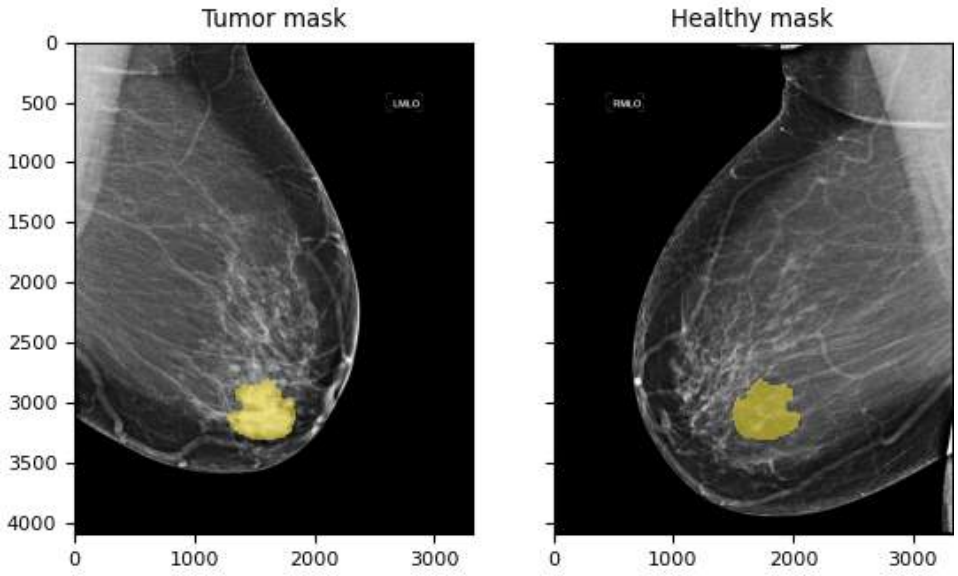


Figure 6.2: Tumor area and area of healthy counterpart

Now in the processed images, the pixel values in these two areas are compared. Figure 6.3 shows the tumor and healthy area after processing, together with a histogram of the pixel values in these areas.

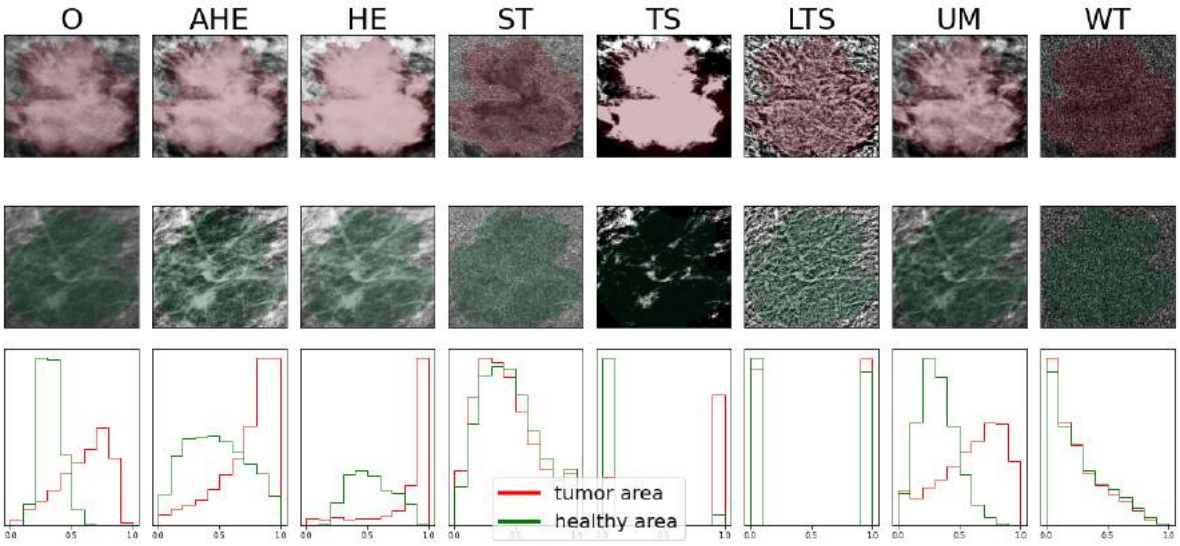


Figure 6.3: Upper row: resulting tumor area after image processing, second row: resulting healthy area after image processing, third row: histogram of pixel values from these areas. From left to right: original image, adaptive histogram equalization, histogram equalization, sobel transform, threshold segmentation, local threshold segmentation, unsharp masking, wavelet transform

Then, the ROC curve and the AUC value are determined. The results for all methods for the same example image is given in Figure 6.4.

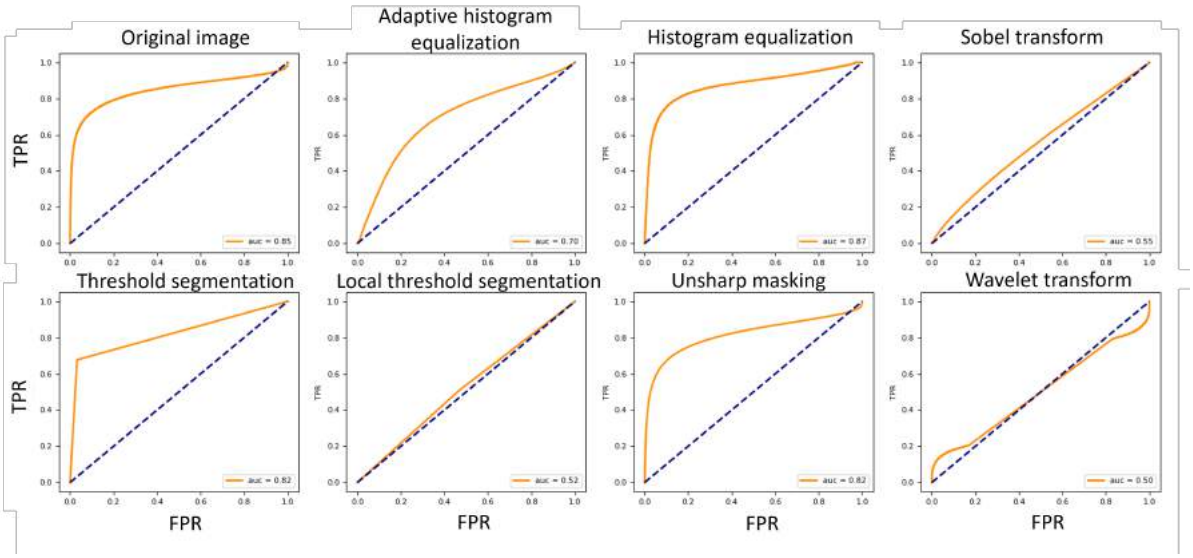


Figure 6.4: Resulting AUC for image processing methods for example patient

From this image the same results follow as the visual inspection. Adaptive histogram equalization, histogram equalization, threshold segmentation en unsharp masking preserve the contrast of the tumor against the background, which follows from the high AUC value. The methods, with lower AUC values, do not preserve this contrast. This figure only shows the results for one image. Therefore, this analysis is done for all images from the EMC dataset, and Figure 6.5 shows the resulting AUC values of all images.

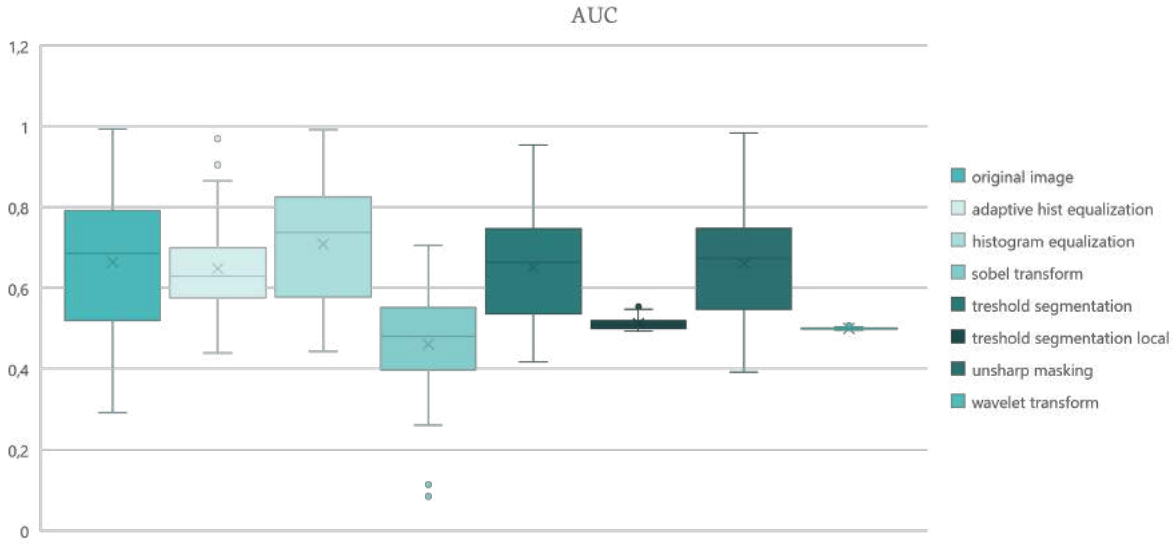


Figure 6.5: Boxplots of AUC values for image processing methods

Histogram equalization is the only method that has a higher average AUC than the original image. The average AUC value of unsharp masking and threshold segmentation is comparable to the original image, while for the other methods the AUC has decreased more. Sobel transform, local threshold segmentation and wavelet transform have a very low average AUC.

6.2. Deep learning

First, the results of existing models for breast cancer detecting are given. Then, the results of transfer learning are given and compared to a model trained from scratch and models trained with preprocessing. The results are compared with the breast density and the opinion of radiologists.

6.2.1. Performance of pre-trained deep learning models

Table 6.1 shows the results of the pre-trained models on the test sets, which are described in subsection 5.2.4.

Model	Original paper	CSAW test set - all	CSAW test set - ILC	EMC set
NYU - image	0.83	0.77	0.65	0.69
NYU - image + heatmap	0.89	0.81	0.72	0.71
GMIC	0.89	0.79	0.66	0.65
GLAM	0.88	0.75	0.64	0.62
End2end	0.88	0.74	0.57	0.59
Faster-rcnn	0.95 (on test set) 0.85 (on challenge set)	0.74	0.62	0.61

Table 6.1: Resulting AUC values of existing deep learning model for breast cancer detection on test datasets

What strikes immediately is that the AUC reported in the papers presenting the models are much higher than the AUC on all three datasets. This is in line with our expectations for the two ILC datasets, since the datasets used in the original studies included images from both IDC and ILC, while our datasets included only images of ILC which is known to be less visible on mammography. Furthermore, the distribution of healthy images and images displaying a tumor are different in our datasets than in the original papers. The second model, *NYU - image + heatmap*, has the best performance results on all datasets. The reason that this model is not selected for transfer learning is that it actually consists of two models, of which one produces a heatmap and a second one produces the classification. Therefore, training is more complicated than in the GMIC model which is chosen for transfer learning. Moreover, the *NYU - image + heatmap* model requires annotations on pixel level, which are not available for all ILC images in the training dataset. GMIC is chosen for transfer learning, since it just requires labels on image level, it is end-to-end trainable, and the performance is still relatively high.

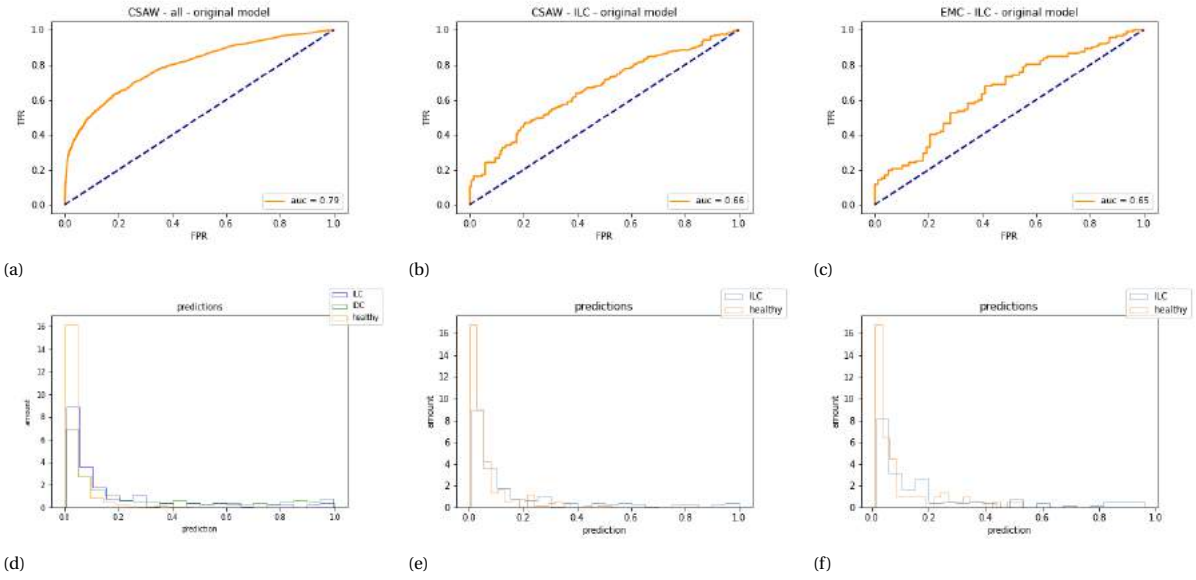


Figure 6.6: ROC curves for original GMIC model on a) CSAW test set including all tumor types, b) CSAW test set with only ILC images and c) EMC set, and d-f) histograms with the tumor predictions, in the same order.

In Figure 6.6 the ROC curves are shown for the GMIC model on the CSAW dataset, for both the set with all

tumor types and the set that is comparable to the EMC dataset, with only ILC data. Also, the ROC curve for the EMC dataset is shown. Furthermore, the histograms for the predictions done by the models are shown in the second row of the figure. These histograms show the distribution of the predictions from the model, split out per healthy images and images containing a tumor. For the dataset containing both ILC and IDC tumors, the predictions are also separated for these groups, as shown in Figure 6.6d. A value close to 0 indicates a low prediction, so a low chance of a tumor present in the image. A higher value, close to 1, indicates a high chance of a tumor present in the image.

From these histograms a few things stand out. Firstly, for images that have the label healthy, the highest peak in prediction values is close to 0, and few images get a high prediction value. This can be observed in all three datasets. For images that have the label tumor, however, the same thing is observed in all datasets. The highest peak in predictions is also close to 0, which indicates that the images are not correctly labeled as tumor. The histogram of the dataset containing all tumor types (Figure 6.6d) also shows a small part of images that received a higher prediction value from the model, so they are correctly classified. The part of IDC images that get a higher prediction is a little bit higher than the part of ILC images that got a higher prediction. This could be an indication that ILC is less well detected by deep learning models than other breast tumor types. Also, the histograms from datasets only containing ILC images in Figure 6.6d and Figure 6.6f show a very small part of images that get a high prediction value.

6.2.2. Transfer learning

Transfer learning on the GMIC model is performed as described in subsection 5.2.3. During training, the loss is evaluated on the training set and on the test set in each epoch. This is shown in Figure 6.7, where each line represents a learning rate, for either the training set (red) or test set (blue). The darker the color of the line, the higher the learning rate that is used. This is the case for both the blue lines with the loss over the test set and the red lines with the loss over the training set. In this figure a few things stand out. First of all, the loss

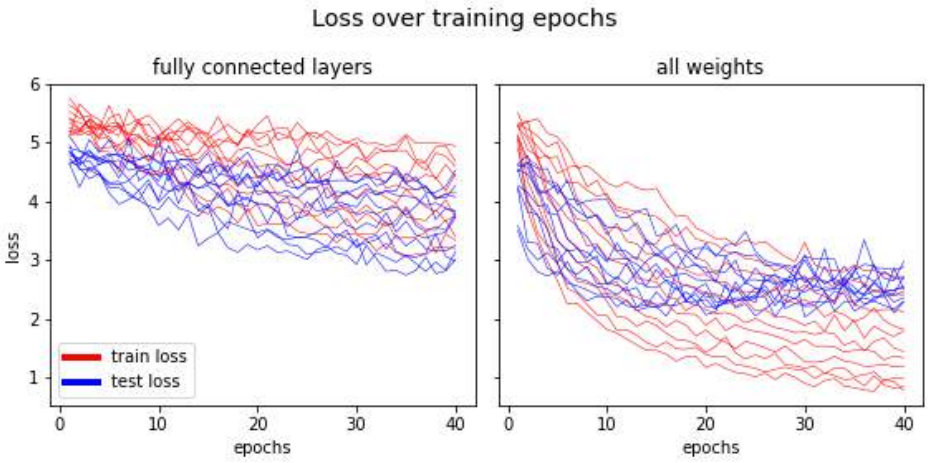


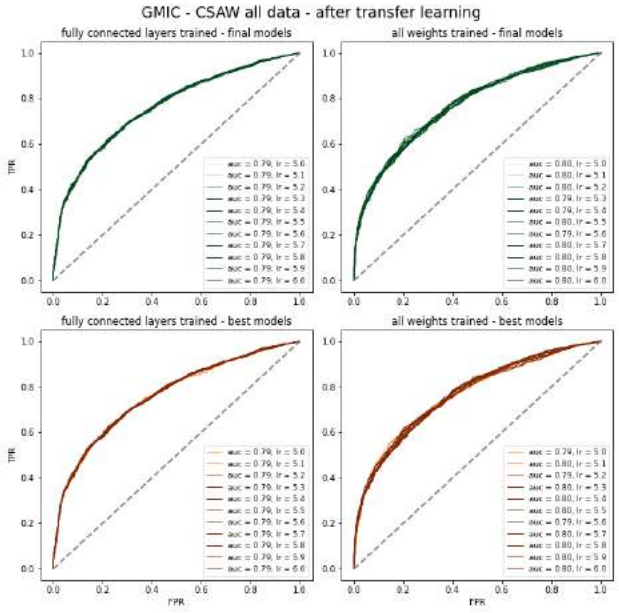
Figure 6.7: Loss over epochs during transfer learning in training set and test set, consisting of healthy and ILC images

for both the training set and test set decrease more during training in the case where all weights in the model are being updated. This indicates that the features that are extracted from the images in the original model are not optimal. In particular, during the first 20 epochs there is a strong decrease in loss for both the test and the training set when all weights are being updated. For the test set, the loss does not decrease so much more after this, but the loss for the training set still decreases. This could be a sign of overfitting. In the case of transfer learning on the fully connected layers, the decrease in loss is less strong. However, there is less sign of overfitting, since the loss for both the test and test set decreases at approximately the same slope during all epochs. Lastly, a higher learning rate results in a steeper decrease of the loss of the training set. For the test set, this result can not be observed.

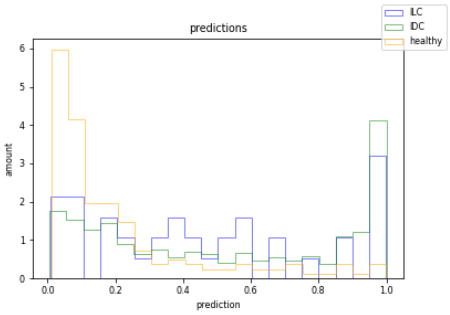
Results of transfer learning models on CSAW data

The models that are created during transfer learning are firstly evaluated on the CSAW dataset, using the images that are not used during training. The resulting ROC curves are shown in Figure 6.8a. Additionally,

for the model with the highest AUC, the histogram of predictions is shown in Figure 6.8b. The resulting AUC



(a) ROC curves of resulting models after transfer learning



(b) Distribution of predictions after transfer learning, best performing model

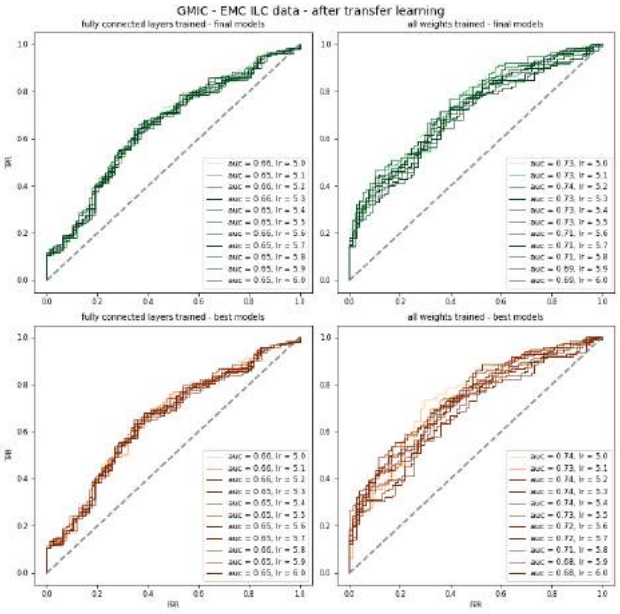
Figure 6.8: Results of transfer learning on all CSAW data

for the models with only the fully connected layers trained, is the same for all different learning rates. This is quite surprising, since Figure 6.7 shows a difference in loss during all training iterations and also in the final model for the different learning rates. What should be mentioned, is that the loss displayed in Figure 6.7 is calculated on the test set containing ILC and healthy images, and the results displayed here are much broader set of healthy images and tumor images from all tumor types. Looking at the resulting AUCs for the models where all weights are trained, the same can be observed. The curves are quite similar and the AUC values lie close together. Again, there is no relation visible between the learning rate and the resulting AUC. Looking at the histogram in Figure 6.8b, some differences can be observed, when comparing this to the histogram belonging to the original model in Figure 6.6d. Still, the highest peak for healthy images is located at values close to 0. However, for both ILC and IDC images, the peak in predictions changed from being close to a very low prediction value, to a high prediction value, close to 1. This indicates that the tumor in these images is now better detected than before transfer learning. In the original model, there were no healthy images that got a prediction value higher than 0.5. This result shows a difference here, with some healthy images getting a prediction higher than 0.5. This means that this model results in more false positive cases than the original model. Comparing the predictions for ILC and IDC in the transfer learned model, the IDC predictions are a bit higher than the ILC predictions, indicating that the model is still able to detect IDC a bit better than ILC.

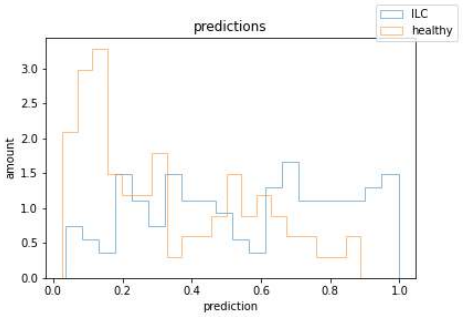
Results of transfer learning models on EMC data

In the same way as the previous section, the results for the transfer learned models are determined on the EMC data, and given in Figure 6.9a. The best performing model is the final model for transfer learning on all rates, with learning rate $10^{-5.2}$, with an AUC value of 0.74. From now on, this model is called the TL (transfer learned) model and used in most of the following analyses. For this TL model, the distribution of the predictions after transfer learning, are determined. They are displayed in Figure 6.9b. Comparing this distribution to the distribution of the predictions from the original model in Figure 6.6f shows some big differences. The peak for ILC images at low prediction values disappeared, and more ILC images get a higher prediction value. The peak for healthy images is still present at the low prediction values, but decreased in hight, since many images also got higher prediction values. Similar to the previous results on the CSAW dataset, this model therefore shows more false positives on the EMC dataset.

To gain more information in what the model did learn specifically during transfer learning the tumor predictions are compared between the original model and the TL model. In Figure 6.10 there are 4 example



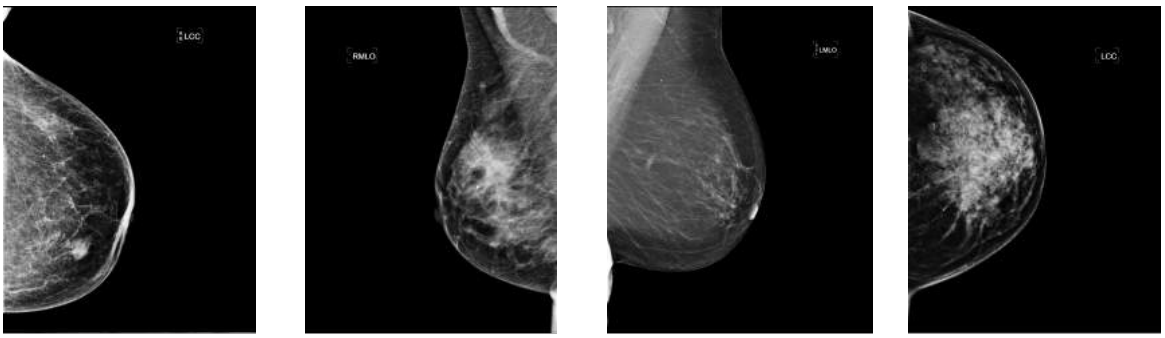
(a) ROC curves of resulting models after transfer learning



(b) Distribution of predictions after transfer learning, best performing model

Figure 6.9: Results of transfer learning on EMC data

mammograms shown that stand out during this comparison, that all contain a tumor. Figure 6.10a shows a mammogram that got a high prediction from both models. Figure 6.10b shows a mammogram that first got a relatively low prediction, but after transfer learning it got a high prediction. The mammogram in Figure 6.10c has the opposite situation, and got a lower prediction score after transfer learning. The last mammogram in Figure 6.10d got a low score from both models. Under each image, the tumor prediction before and after transfer learning is shown, together with the percentage of healthy images that got a lower prediction score than this image. This percentage is added, since the distributions of predictions are different in both models and therefore the two predictions for one image can not directly be compared to each other. In the ideal situation, 100% of healthy images should retrieve a score, lower than this mammogram which contains a tumor.



(a) (b) (c) (d)

Figure 6.10: Example mammograms from EMC dataset with different results from transfer learning a) mammogram with tumor detected by both models b) mammogram with tumor not detected in original model, detected in TL model c) mammogram with tumor detected in original model, not detected in TL model d) mammogram with tumor not detected by both models. Under each image, the tumor prediction before and after transfer learning is shown, together with the percentage of healthy images that got a lower tumor prediction score.

The images that were correctly classified by the original model (Figure 6.10a and Figure 6.10c) both show

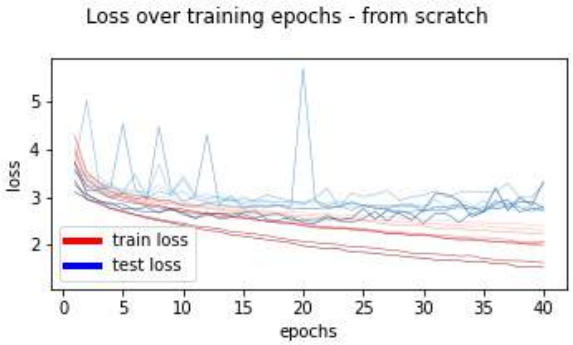


Figure 6.11: Value of loss function over epochs during training from scratch in training set and test set

a white mass, that stands out against the darker background. The mass in Figure 6.10c could be too small to be detected and correctly classified by the TL model. The images that were wrongly classified by the original model (Figure 6.10b and Figure 6.10d) show more dense tissue, and therefore the tumor is less well visible. In Figure 6.10b the white area is brighter and bigger and stands out a little bit against the heterogeneous background, which could be a reason for the TL model to detect it. The tumor tissue in Figure 6.10d just does not stand out as much against the background tissue, which could be the reason for both models to not classify this image as tumorous.

6.2.3. Model training from scratch

To determine the added value of transfer learning in these results, models are trained where the weights from the original model are not used as initialization. These models are validated on both the test part of the CSAW dataset as well as the EMC dataset. The resulting AUC values on these datasets are much lower than the original model, or the model that the transfer learning was applied to, for both methods of transfer learning. The AUC values on the CSAW test set range between 0.44 and 0.57, compared to 0.80 for the TL model. For the EMC set the range of AUC values from the models from scratch is 0.46 - 0.60, compared to 0.74 for the TL model. Therefore, the added value of transfer learning and initializing the model before training with the weights from the original study is substantive. To further investigate why these results are lower than the TL model, Figure 6.11 shows the value of the loss function over the epochs. Again, the darker lines represent higher learning rates. Comparing these result to Figure 6.7, the loss for the test set is much less stable than the model with transfer learning, with big outliers up. The loss for the training set keeps decreasing, while the loss for the test set does not decrease anymore and sometimes even increases. This is a sign of overfitting of the model on the training set.

6.2.4. Transfer learning with pre processing

Transfer learning with pre processing is conducted, and the results on the AUC values when these models are tested on the different datasets is given in Table 6.2. To enable comparison, the results of the original model and the transfer learned model are also given. The range of the AUC values for different learning rates is given, because the models differed quite much in the resulting AUC.

Preprocessing	**CSAW test set - all**	**CSAW test set - ILC**	**EMC set**
Original model	0.79	0.66	0.65
Transfer learning without preprocessing	0.79 - 0.80	0.68 - 0.76	0.65 - 0.74
Histogram equalization	0.67 - 0.70	0.63 - 0.69	0.61 - 0.66
Unsharp masking	0.58 - 0.64	0.57 - 0.66	0.55 - 0.63
Sobel transform	0.54 - 0.55	0.51 - 0.61	0.49 - 0.57

Table 6.2: Range of resulting AUC values for models trained using pre processing methods

Especially in the smaller datasets *CSAW data - ILC* and *EMC data*, the range in the AUCs is big. Therefore, it might be the case that some high AUC values are coincidental. For histogram equalization, the results for the *CSAW data - ILC* and *EMC data* are sometimes higher than the original model. However, transfer learn-

ing without preprocessing still results in higher AUC values for both these datasets. Transfer learning with unsharp masking and sobel transform result in AUC values that are even lower than the original model.

6.2.5. Spatial analysis

The heatmaps are compared with the segmentation by the breast radiologist. This is done for the original model and the best performing transfer learned model. In Figure 6.12 the heatmaps are shown for the example images from Figure 6.10, together with the segmentation by the breast radiologist.

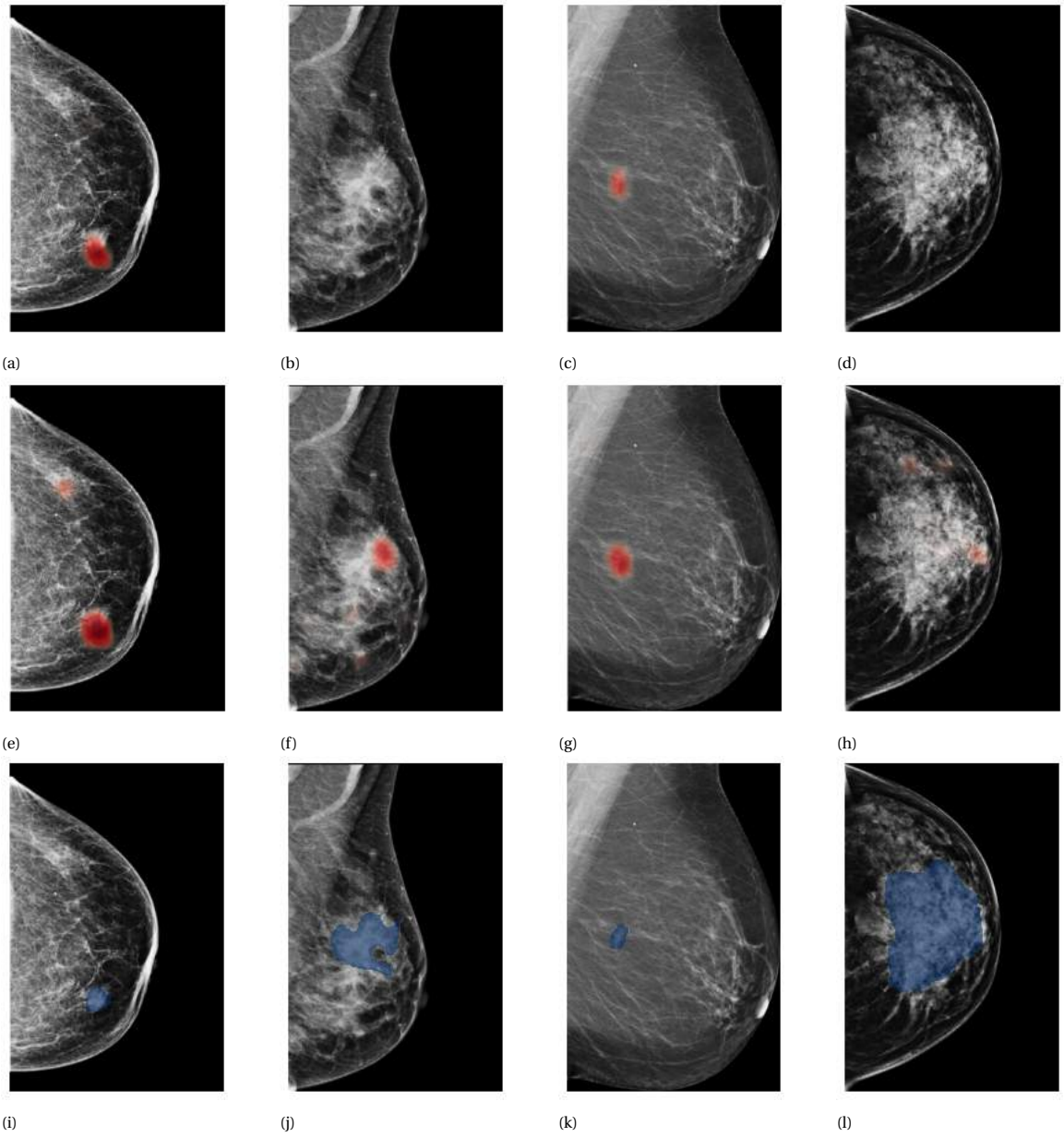


Figure 6.12: Resulting heatmaps for mammograms for example images. a-d) original model, e-h) new model, i-l) segmentation by breast radiologist

Following the steps in subsection 5.3.1, Dice's coefficient is calculated for each image, and the range of resulting Dice's coefficients is shown in Figure 6.13 for the original and TL model.

When the maximum value of the heatmap is taken, and compared to the true segmentation, the results show that in the original model 37% of the heatmap has its maximum value in the true segmentation, compared to 58% in the TL model.

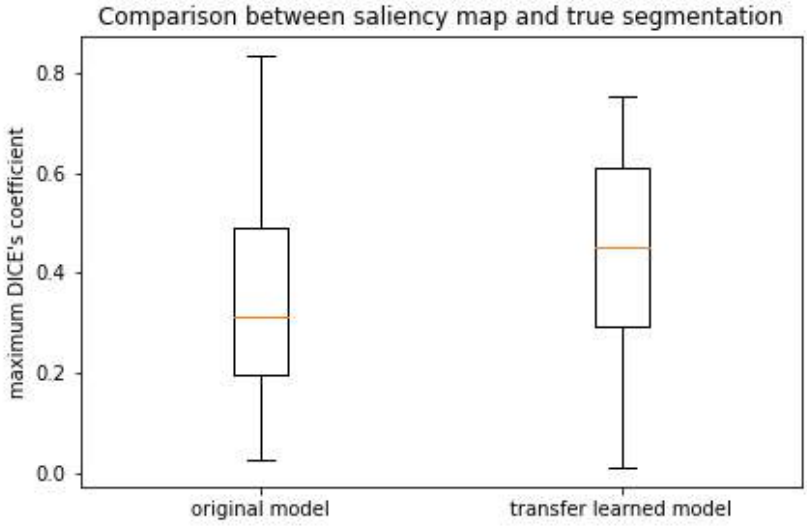
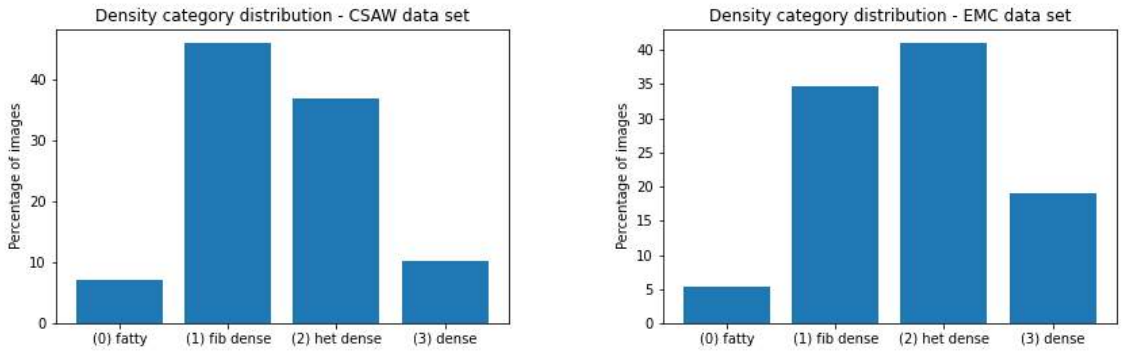


Figure 6.13: Comparison of DICE's coefficients for original and TL model

6.2.6. Comparison of performance by different breast densities

Using the prediction model of Wu et al. (2018), the images are split into a dense (category 0 and 1) and a non-dense (category 2 and 3) category, and the results are determined for both groups. Like described in section 1.4, the four categories are (0) Almost entirely fatty, (1) Scattered fibroglandular density, (2) Heterogeneously dense and (3) Extremely dense. Figure 6.14 shows the distribution of images from the *CSAW test set - all* and the *EMC set* that belong to the four different density categories. Most images belong to categories 1 and 2, and a smaller part of the images belongs to the two more extreme categories 0 and 3. For the *EMC set*, a higher percentage of images belongs to the highest density category than for the *CSAW test set - all*. Table 6.3 shows the resulting AUC values for the dense and non-dense category, for the CSAW dataset and the



(a) CSAW test set - all

(b) EMC dataset

Figure 6.14: Distribution of density categories (0) Almost entirely fatty, (1) Scattered fibroglandular density, (2) Heterogeneously dense and (3) Extremely dense,

EMC dataset. From this table it can easily be determined that the AUC value is higher for images from the

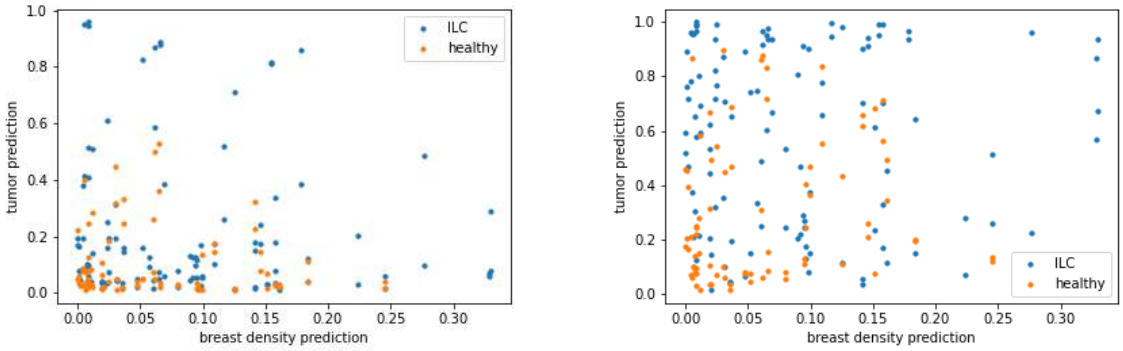
non-dense categories, for all datasets. This is observed both in the original model as well as the model after transfer learning.

		AUC dense (category 2 and 3)	AUC non-dense (category 0 and 1)	AUC without most dense group (category 0, 1 and 2)
Original model	CSAW test set - all	0.74	0.83	0.80
	CSAW test set - ILC	0.60	0.70	0.68
	EMC set	0.61	0.75	0.65
TL model	CSAW - all	0.74	0.84	0.77
	CSAW- ILC	0.66	0.75	0.74
	EMC set	0.70	0.83	0.75

Table 6.3: Results of original model and TL model, split in dense and non-dense breasts

Secondly, the most dense category is left out of the analysis. Like described in section 5.3, this is to simulate the effect of additional MRI screening for this group of women with breast density from the highest category. These results are also shown in Table 6.3. The AUC values of this group lie in between the values of the two other groups, as can be expected.

To gain a better understanding into the predictions for the different breast density categories, the prediction value is plotted against the breast density value. This value is acquired from the breast density model, as the prediction that the image belongs to the most dense category. In this way, the highest density value belongs to the highest breast density. These results are visualized for the EMC dataset in Figure 6.15. In Figure 6.15a this is done for the original model, and in Figure 6.15b the same is shown after transfer learning.



(a) Original model

(b) TL model

Figure 6.15: Comparison of breast density to tumor prediction for EMC data

The most outstanding from the first figure is that for images that have a high density value, the tumor predictions is low. This indicates that the model cannot detect the tumor when the breast tissue is dense. The highest tumor predictions are from images with a low breast density. In the second figure there is a different pattern, since images from all breast density values get high tumor predictions.

6.2.7. Comparison to radiologists

Firstly, the labels given by the radiologists are considered, and compared to the true labels. This is shown in Table 6.4 and Table 6.5, both for the *CSAW test set - ILC* and the *CSAW test set - all*.

From these tables, the sensitivity and specificity are calculated. The sensitivity of the radiologists on the *CSAW test set - ILC* is 78% and the specificity is 98%. For the *CSAW test set - all* the sensitivity is 59% and the sensitivity is also 98%. The sensitivity for the *CSAW test set - all* is low, compared to previous studies on breast cancer detection from mammography. A possible cause for this is the nature of the dataset and its labels. A mammography image received a tumor label, when in the time following on the screening a tumor

		Radiologist label	
		1 (tumor)	0 (healthy)
True label	1 (tumor)	28	8
	0 (healthy)	78	3702

Table 6.4: CSAW test set - ILC

		Radiologist label	
		1 (tumor)	0 (healthy)
True label	1 (tumor)	196	138
	0 (healthy)	78	3978

Table 6.5: CSAW test set - all

Confusion matrices to compare the labels given by radiologists to the true labels

developed. This time in between screening and tumor development could be a maximum of 729 days, which is the time between two screenings. Therefore, maybe just a small beginning of the tumor was present at the time of screening. Another thing that stands out is the high sensitivity in the *CSAW test set - ILC*, compared to the *CSAW test set - all*. A cause for this difference could be the small size of the dataset, which enables coincidental outcomes.

To compare these results to the detection models, a threshold value must be chosen to transform the tumor prediction values into a classification into tumor and healthy. The threshold value that is used, is the threshold that makes the model perform comparable to healthy images to the radiologists. This means that the sensitivity of the model equals the sensitivity from the radiologists of 98%. This is done both for the original GMIC model and the TL model, for both datasets.

The thresholds that follow from this method are 0.39 for the original model and 0.89 for the TL model. Using these thresholds, Table 6.6, Table 6.7, Table 6.8 and Table 6.9 are generated.

		Model label	
		1 (tumor)	0 (healthy)
True label	1 (tumor)	6	30
	0 (healthy)	72	3708

Table 6.6: Original model, CSAW test set - ILC

		Model label	
		1 (tumor)	0 (healthy)
True label	1 (tumor)	89	245
	0 (healthy)	99	3957

Table 6.7: Original model, CSAW test set - all

		Model label	
		1 (tumor)	0 (healthy)
True label	1 (tumor)	7	29
	0 (healthy)	74	3706

Table 6.8: TL model, CSAW test set - ILC

		Model label	
		1 (tumor)	0 (healthy)
True label	1 (tumor)	81	253
	0 (healthy)	87	3969

Table 6.9: TL model, CSAW test set - all

Confusion matrices to compare the labels given by the GMIC detection models to the true labels, CSAW dataset, specificity = 98%

The resulting sensitivities are shown in Table 6.10. These values seem low, compared to the radiologists. However, the thresholds can be chosen differently, when a lower specificity would be allowed. To investigate the result of this, the same calculation for sensitivity is made for thresholds that leads to a specificity of 90%, and these results are also added to Table 6.10.

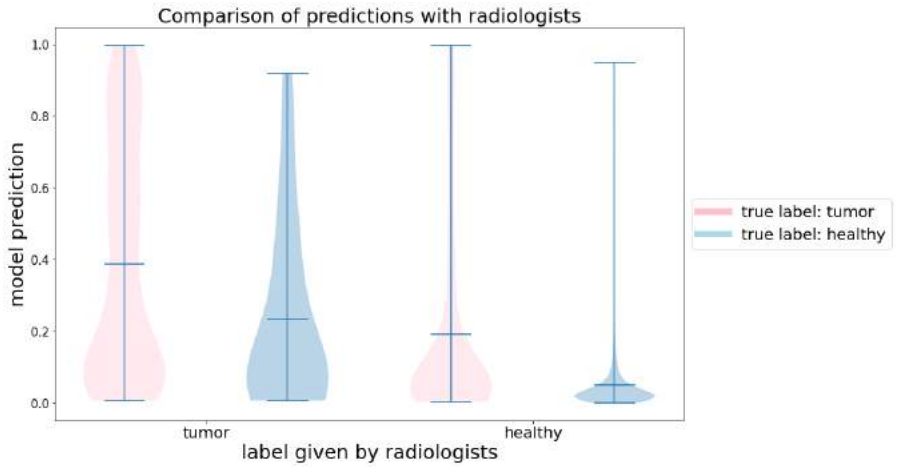
		Dataset	
		CSAW test set - ILC	CSAW test set - all
Original model	Specificity 98 %	17%	27%
	Specificity 90 %	44%	46%
TL model	Specificity 98 %	18%	24%
	Specificity 90 %	47%	40%

Table 6.10: Sensitivities of original and TL model for *CSAW test set - ILC* and *CSAW test set - all* for the specificity values of 98% and 90%

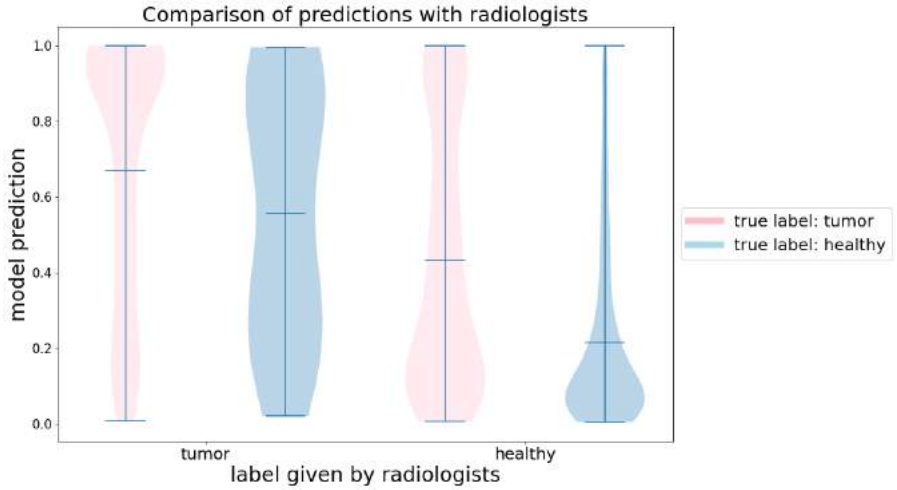
For further comparison of the models and the radiologists, Figure 6.16 is added, combining the model outcome to the radiologist reviews. This comparison is done for the *CSAW test set - all*, because the number of tumors in the *CSAW test set - ILC* is too low to detect a distribution in tumor prediction values. Figure 6.16b

shows the results for the original model and **??** for the TL model. The vertical axes represents the outcomes of the model, thus the prediction of the presence of a tumor, by the models. The four areas that are shown, each represent a different group of images, which are the images from the four categories in **??**. The two bars on the left side show images that got labeled as tumor by the radiologists. The pink bar shows the results for the subset of these images that actually contained a tumor. Therefore, the True Positive group. The second bar, shows the results for the False Positive group. The third bar the False Negative group, and the fourth bar the True Negative group. The small horizontal line in the bar indicates the average tumor prediction value that is given to the images in the group by the model.

Looking at these figures, some observations can be made. Firstly, looking at the images that were labeled as tumor by radiologists (left two bars), the images that actually contained a tumor got an average prediction of ± 0.4, while the images that were indeed healthy got a lower average prediction of ± 0.25. Therefore, the model could assist in determining whether an image that is labeled as tumor, actually contains an image. The same holds for the images that were labeled as healthy by the radiologists. Of these images, the ones that actually contained a tumor got an average tumor prediction value of ± 0.2, while the images that were indeed healthy got an average prediction value of ± 0.05. Applying the model to the images that the radiologists labeled as healthy, could therefore help in finding tumors that are missed. For the TL model, similar patterns are observed.



(a) Original model



(b) Model after transfer learning

Figure 6.16: Result of model predictions, comparison with radiologists, a) original model, b) model after transfer learning

To truly understand the added value of these two models to the screening by radiologists, the most important

images are the ones that actually contained a tumor, but were labeled by the radiologists as healthy. These tumors were missed and could have been earlier detected. Therefore, Table 6.11 shows all images that were labeled as healthy by the radiologists, and the outcome of the two models using the threshold as before for a specificity of 90%. These results show for the original model, that if a specificity of 89% is allowed, 23% of the tumors that were not detected by the radiologists, would have been detected by the model. For the TL model, if a specificity of 90% is allowed, 20% of the tumors that were not detected by the radiologists, would have been detected by the TL model.

	Sensitivity	Specificity
Original model	23%	89%
TL model	20%	90%

Table 6.11: Performance of models on images that are classified as healthy by radiologists

7

Discussion

7.1. Discussion of results

The classical image processing techniques were used as an exploratory research into the properties of mammography images. The methods that are applied turn the input image into an output image, with different properties. What the results mostly show, is that image processing techniques alone do not have significantly different effects on tumor area and healthy area. Therefore, these methods do not contribute to the task of distinguishing a tumor area from the background. However, the method that is used to evaluate this is the AUC value. This value depends on the histogram of the pixel values in the healthy and tumor area, and quantifies how far apart these histograms are from each other. Consequently, the structure or patterns visible in these areas are not taken into account. Other evaluation methods that do take structure into account could give different insights in the function of the image processing methods. Moreover, combinations of image processing methods could be applied. Lastly, there are methods that are not tested. The decision to not take all these steps to improve the image processing methods lies in the focus of this study. These decisions are made, based on previous studies, showing that deep learning has a higher potential in mammography evaluation, which is why the focus in this research was on deep learning.

Concerning the trained deep learning models that are tested, the AUCs that are found for the datasets that are used in this study are lower than presented in the original papers (K. Liu et al., 2021; Ribli et al., 2018; L. Shen et al., 2019; Y. Shen et al., 2019; Wu et al., 2020). This holds for both the EMC and the CSAW dataset. For the CSAW dataset, containing all tumor types and many healthy mammograms as well, the differences between the AUC in the original papers describing the models and the AUC found in this study are bigger than expected. A possible cause for this could be differences in the properties of the dataset, or different quality of the mammograms.

After transfer learning on the GMIC model, the tumor prediction values of ILC images increase, as it is intended. Also, the tumor prediction values of IDC increase, which indicates similarities in the features of IDC and ILC that are visible on mammography. Interestingly, also the tumor prediction values of healthy images increase substantially. A reason for this increase could be that the model is now focused on the detection of ILC, which is less visible, and therefore the model becomes more sensitive and is more inclined to give higher predictions. The shift from low to high tumor prediction values raises the question of what the best threshold value would be to define the image as tumor. For the usual threshold value of 0.5, the original model would result in many false-negatives, but almost no false-positives. Therefore, an argument could be made to lower the threshold to decrease the number of false-negatives of missed tumors, but this would also result in more false-positives. In the transfer learned model, the threshold value of 0.5 would approximately detect half of the tumors, with a relatively low number of false-positives. Again, when this threshold is lowered, more tumors will be detected, also raising the number of false-positives. These considerations should be taken into account, when a threshold is decided on.

For a more general dataset, consisting of all tumor types, the increase in performance of the model is lower than for a dataset with ILC images. This indicates that training with ILC images does learn the model features

that are specific for ILC.

The initialization of the weights in the model before transfer learning with the weights from the original model is very important. When the model is randomly initiated, the performance of the model after training is much worse than with the initialization form the original model. This result is as expected, since the 152 images containing ILC that are used in training are not enough to train all the weights in this complicated network. The original network, trained with 186.816 images of which 1714 had a malignant finding, already incorporated many of the features that are essential for breast cancer detection. Efficiently using multiple datasets consisting of many images has a high potential to improve the performance of breast cancer detection models.

From the results regarding the performance of the models for the different categories, it can be observed that high breast density influences the detection performance. This result is as expected and it matches the performance of radiologists. Like mentioned before, dense breast tissue covers up the tumor tissue and in this way the tumor becomes less visible. Therefore, this result is as expected.

Considering the spatial results of the model, some points of discussion should be mentioned, that decrease the accuracy of the predicted segmentations. The first goal of the saliency maps is not to segment the tumor perfectly or give the exact location of the tumor, but instead it enables the model to determine the approximate areas of interest to use as input for the local module. Furthermore, the radiologist segmented the tumor area as well as possible, but this true segmentation cannot be expected to be perfect. Therefore, the resulting DICE's coefficients are not expected to be precise, but rather to give an indication of the correctness of the saliency maps. The same holds for the comparison of the maximum value of the saliency maps to the true segmentation. These results both show that the saliency maps cover the tumor area in most cases, but definitely not find the correct area in all images. A cause for this is the weakly supervised approach, in which image labels are used to determine the saliency maps on pixel-level. Deep learning using true segmentations will probably results in better spatial results.

Lastly, the model outcomes are compared to radiologist reviews. This comparison indicates that in general, images that are correctly classified by the models are also correctly classified by the radiologists. Furthermore, to obtain similar levels of specificity as radiologists, the sensitivity of the models are lower than radiologists. However, in the images that are classified as healthy by the radiologists, the models succeed in the detection of some tumor images, reducing the amount of missed tumors during screening.

7.2. Limitations

The first limitation for this study follows from the data sources. From the Erasmus MC dataset, consisting of images from 162 patients with confirmed ILC, only 55 patients were used for evaluation of the methods. This is due to the available time investment of the radiologist to mark the images. When all images are marked by a radiologist, a more extensive evaluation of the methods can be done. The results suggest that retraining a model with only ILC images, improves the performance of the model on other ILC images. However, with a small test size, caution must be applied, as the findings might not be applicable to all ILC images. The CSAW dataset that is used for training of the model, consists of more images with a confirmed tumor. But, since only the ILC images are used for training, this amount is still rather small compared to training datasets used in other studies.

Another issue that should be discussed is the availability of models that already have been trained with a dataset. The models that are compared in this study, are publicly available. However, the majority of trained models that has been developed is not publicly available. Some models are used in the development of commercial products, which is a reason for not making them available. As a result of this, the comparison in this study is not complete, and adding the not-publicly available models to the comparison would be a great addition to this field of study.

7.3. Recommendations

7.3.1. Future research

The first recommendation for future research is to include more models in the evaluation of current methods. Like stated before, a significant share of developed models is not publicly available. However, more effort could be put into acquiring access to these models for the purpose of evaluating and comparing them on the same datasets.

In terms of the structure of the model, the recommendation is to evaluate the added value of using pixel level segmentations in the training. In the weakly supervised approach in this study, pixel level segmentations are only used during evaluation and not during training. The advantage of this method is that more extensive datasets can be used for training, that are not annotated on pixel-level. However, when the contribution of using these annotations during training is proven to be significantly high, it could be preferable to focus on acquiring more pixel-level annotations, instead of focusing on developing weakly supervised training methods.

Studying the results, using different distributions of tumor types in the training data is another recommendation regarding model training. Now solely ILC data has been used for training, but this also resulted in a limited amount of images used. Adding images with other tumor histology types would result in a bigger set of training data, which could potentially improve the performance of the model.

Lastly, on the clinical application of breast cancer detection models, future research would also be recommended. There are multiple options for the application of these models. Moreover, there a many ways in which a radiologist can work together with a prediction model. A model could firstly scan all images, and preselect images of which it suspects that an abnormality is present. These images would then be shown to a radiologist, which makes the final decision. Another option is to use the model as a second reader, and use only one human reader. A third option is to use the model on all images, and present the results of the model together with the original images to the radiologists, and let the screening continue the same as now. A last option could be to keep the screening the same as now, but apply the model on all images that are classified as healthy by the radiologists. Then, for hard cases where humans did not find a tumor, maybe the model can in some cases detect it. Research into these options and the effects of them on the clinical practice is needed.

7.3.2. Data collection

In this study, images are collected from the Erasmus MC database. However, not all images could be segmented, due to time restrictions. AT the Erasmus MC, future research in the detection of ILC and other tumors would benefit from the additional labeling and segmentation of these images. Moreover, in the Erasmus MC patients with all tumor types are treated. The dataset could therefore be expanded with these images, to enable research that does not only focus on ILC. Furthermore, there is ongoing research on federated learning for cancer detection, like the EuCanImage initiative for example. A more elaborate data collection from the Erasmus MC could be a value to these initiatives. Another option for data in future research is a collaboration with the dutch population screening programme, where all women are screened from the age of 50 years. This could be a big source for mammography images.

8

Conclusion

In the Netherlands, approximately 15.000 women are diagnosed with breast cancer yearly. Population Screening Programs worldwide have the intention to detect breast cancer in an early stage and therefore provide the possibility of an appropriate treatment timely. Because radiologists sometimes do not detect breast cancers, automatic breast cancer detection is a growing field of research. Especially, ILC is often missed during the population screening, since it is less visible on mammography. To recap, the main research question of this project is:

Can image processing techniques and deep learning models contribute to the earlier detection of Invasive Lobular Carcinoma from mammograms?

The results of the image processing techniques show that the basic operations that can be performed on images, do not contribute much to the visibility of ILC on mammograms. Histogram equalization is the only method with an average AUC that is higher than the original model, when the pixel values of a tumor area are compared to the pixel values of a corresponding healthy area. Unsharp masking is the method that got the second highest average AUC, but this was lower than the original image.

Existing deep learning models that are publicly available, have a lot of potential for mammography classification. The maximum AUC value that is found for the CSAW dataset is 0.81 and for the EMC dataset, consisting only of ILC images, the maximum AUC value is 0.71. For the GMIC model, these values are 0.79 for the CSAW dataset and 0.65 for the EMC dataset. Transfer learning with ILC and healthy images improved these values to 0.80 for the CSAW dataset and 0.74 for the EMC dataset. This shows that for ILC images specifically, transfer learning is a worthwhile addition.

The results in this study show clearly that for images of breasts with a high breast density, the tumor prediction works less well than for images of breasts with low breast density. This is also observed when breast radiologists evaluate the images.

In general, there is a high resemblance between the reviews of radiologists and the outcome of the GMIC model, both before and after transfer learning. Images that were wrongly classified by the models, were often also wrongly classified by the radiologists that reviewed the image. However, there were also images actually containing a tumor that were classified by the radiologists as healthy, while the model classified the image as tumor. This depends on the required specificity for the model. When a specificity of 98% could be allowed, 23% of the missed tumors could have been detected by the original GMIC model. This promising finding shows that the model - after modification and validation - could be of great additional value to the clinical practice, by contributing to early detection from mammograms.

Previous studies already showed the potential of applying deep learning methods to breast cancer detection from mammograms. This study contributes to this field, focusing on the possibilities for the detection of ILC. Deep learning models are capable of ILC detection at the same level as radiologists. Training with ILC images specifically improves this detection. Therefore, it can be concluded that deep learning models can

contribute to the earlier detection of Invasive Lobular Carcinoma from mammograms.

A

Algorithms

A.1. Retrieve region of interest from saliency map

Algorithm 2 retrieve_roi

Input: $\mathbf{x} \in \mathbb{R}^{H,W}, \mathbb{A} \in \mathbb{R}^{h,w,2}, K$
Output $O = \{\tilde{\mathbf{x}}_k | \tilde{\mathbf{x}}_k \in \mathbb{R}^{h_c,w_c}\}$

$O = \emptyset$
for each class $c \in$ {benign, malignant} **do**
    $\tilde{\mathbf{A}}^c$ = min-max-normalization($\mathbf{A}^c$)
end for
$\mathbf{A}^* = \sum_C \tilde{\mathbf{A}}^c$
l denotes an arbitrary $h_c \frac{h}{H} \times w_c \frac{w}{W}$ rectangular patch on $\mathbf{A}^*$
criterion$(l, \mathbf{A}^*) = \sum_{(i,j)\in l} \mathbf{A}^*[i,j]$
for each 1,2,...,K **do**
    $l^* = \text{argmax}_l \text{criterion}(l, \mathbf{A}^*)$
    L = position of l^* in $\mathbf{X}$
    $O = O \cup \{L\}$
    $\forall (i,j) \in l^*$, set $\mathbf{A}^*[i,j] = 0$
end for
return O

A.2. Transform radiologist drawing to tumor segmentation

The following operations are used in the algorithm below, together with its definition and abbreviation: Also,

Operation	Abbreviation	Definition
Binary dilation	BD(A, B)	$A \oplus B = A \bigcup_{b \in B} A_b$ with A_b the translation of A by b
Binary erosion	BE(A, B)	$A \ominus B = \{z \in E \mid B_z \subseteq A\}$, with $B_z = \{b + z \mid b \in B\}, \{\forall z \in E\}$ and E a euclidean space
Binary fill holes	BFH(A)	1. BFH(A) = A 2. $\forall x_i$ with $A(x_i) = 0 \rightarrow \text{BFH}(x_i) = \begin{cases} 1 \text{ if } x \text{ is within an area surrounded by 1's in A} \\ 0 \text{ if x is not within an area surrounded by 1's} \end{cases}$
Convex hull	CH(A)	1. CH(A) = A 2. $\text{conv}\, X = \left\{ \sum_{i=1}^{n} \alpha_i \cdot x_i \,\middle|\, A(x_i) = 1, n \in \mathbb{N}, \sum_{i=1}^{n} \alpha_i = 1, \alpha_i \geq 0 \right\}$ 3. $\forall x_i \in \text{conv}X \rightarrow CH(x_i) = 1$.

Table A.1: Morphological operations used and definitions

masks are used in the shape of ellipses. Ellipse(50) means an ellipse with a height and width of 50. The value for the coordinates inside this ellipse is 1.

Algorithm 3 transform drawing to segmentation

Input: $\mathbf{x} \in \mathbb{R}^{H,W}$ with $x_{i,j} \in \{0, 1\}$
Output: $\mathbf{x_{filled}} \in \mathbb{R}^{H,W}$ with $x_{filled_{i,j}} \in \{0, 1\}$

$x_{dilation_1} = \text{BD}\left(x, \text{Ellipse}(50)\right)$
$x_{erosion_1} = \text{BE}\left(x_{dilation_1}, \text{Ellipse}(50)\right)$
$x_{filled_1} = \text{BFH}\left(x_{erosion_1}\right)$
if $x_{filled_1} = x_{erosion_1}$ **then**
    $x_{dilation_2} = \text{BD}\left(x_{erosion_1}, \text{Ellipse}(75)\right)$
    $x_{erosion_2} = \text{BE}\left(x_{dilation_2}, \text{Ellipse}(75)\right)$
    $x_{filled_2} = \text{BFH}\left(x_{erosion_2}\right)$
    if $x_{filled_2} = x_{erosion_2}$ **then**
        $x_{dilation_3} = \text{BD}\left(x_{erosion_2}, \text{Ellipse}(100)\right)$
        $x_{erosion_3} = \text{BE}\left(x_{dilation_3}, \text{Ellipse}(100)\right)$
        $x_{filled_3} = \text{BFH}\left(x_{erosion_3}\right)$
        if $x_{filled_3} = x_{erosion_3}$ **then**
            $x_{filled} = \text{CH}(x_{erosion_3})$
        else
            $x_{filled} = x_{filled_3}$
        end if
    else
        $x_{filled} = x_{filled_2}$
    end if
else
    $x_{filled} = x_{filled_1}$
end if
return x_{filled}

B

Architectures of trained deep learning models for breast cancer detection

B.1. Description

This section describes the four model architectures and ideas behind it, that are not yet explained in the main text.

B.1.1. Deep Neural Networks Improve Radiologists' Performance in Breast Cancer Screening (Wu et al., 2020)

There are two models described in this article, that are both tested in this research. The first model takes the four views as input and uses a seperate ResNet architecture to extract features for these views. Then, the resulting features from the two MLO images and the features from the two CC images are concenated, and followed by fully connected layers, which result in a classification for four categories: *left benign*, *right benign*, *left malignant* and *right malignant*. These predictions are then averaged over the views. This model is referred to as the *image-only* model. The second model uses the same architecture for the classification of images, but adds two input images per view, which are a result of a patch classifier. This patch classifier takes patches of 256 x 256 pixels, and outputs a prediction for this patch. This classifier is applied to the test images in a sliding window to the full resolution mammogram, to create heatmaps for both benign and malignant findings. These heatmaps are then used as extra input channels for the classification architecture. This model is referred to as the *image-and-heatmaps model*. The architecture of the classification model is shown in Figure B.1.

B.1.2. Weakly-supervised High-resolution Segmentation of Mammography Images for Breast Cancer Diagnosis (K. Liu et al., 2021)

This architecture consists of a global and a local module, like shown in Figure B.2 and is referred to as *GLAM*. The input image is first processed by the global module, which is similar to a ResNet architecture. This global module produces heatmaps, which in turn result in patches with ROI's. These patches are then extracted from the original model with full resolution, and applied to the local module. This results in a saliency map for each patch, which are mapped back to the original image to obtain the final saliency map for the whole image. The classification output is obtained from the global module of GLAM.

B.1.3. An interpretable classifier for high-resolution breast cancer screening images utilizing weakly supervised localization (Y. Shen et al., 2019)

This model, which is similar to *GLAM*, also consists of a local and global module with the same functions. An addition is the fusion module, like shown in Figure 5.2. In this fusion model, features from the global module and local module are concatenated and fed to a fully connected layer to produce the final prediction for a mammogram.

B.1.4. Deep Learning to Improve Breast Cancer Detection on Screening Mammography (L. Shen et al., 2019)

In the study by L. Shen et al. (2019), a patch classifier is transformed to a whole image classifier. The patch classifier has five possible outcome classes, being benign calcification, malignant calcification, benign mass, malignant mass and background. This patch classifier is used as the first layer of the whole image classifier. The five resulting heatmaps, corresponding to the five classes, are fed into top layers, consisting of two blocks of convolutional layers, a global average pooling layer and the final classification into cancer and normal. This structure brings the advantage of the possibility for training with image-level labels. The patch classifier is firstly trained with labels on patch level, which is used as initialization of the whole network. After that, image-labels are used to finetune the network. The network is shown in Figure B.3.

B.1.5. Detecting and classifying lesions in mammograms with Deep Learning (Ribli et al., 2018)

The last breast cancer classification model is based on Faster R-CNN. Faster R-CNN starts with convolutional layers (VGG16), with a Region Proposal Network added on top of these. This RPN detects and localizes objects in the image, and sends the bounding boxes with the highest probability score to the second part of the network. This second part aims to classify the objects in these bounding boxes. The possible classes for this network are benign, malignant and normal. The final score for one image is the maximum probability score for the malignant class for a detected bounding box.

B.2. Visualization

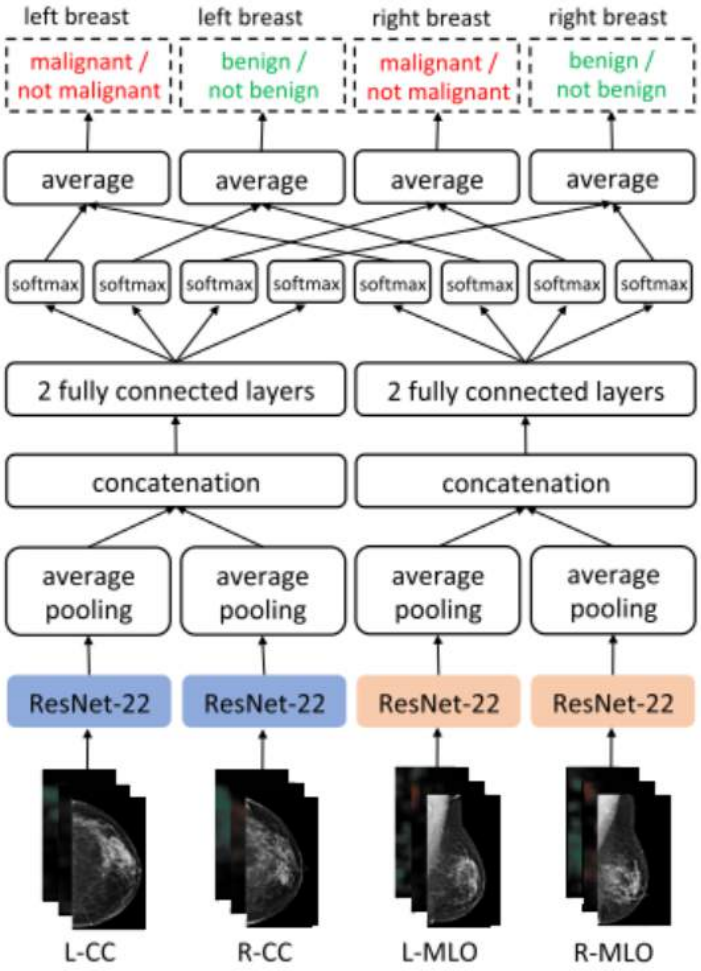


Figure B.1: The architecture of the classification models *image-only* and *image-and-heatmaps* (Wu et al., 2020)

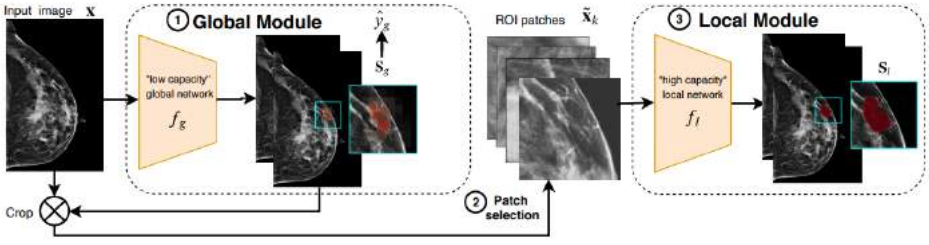


Figure B.2: The architecture of the GLAM model (K. Liu et al., 2021)

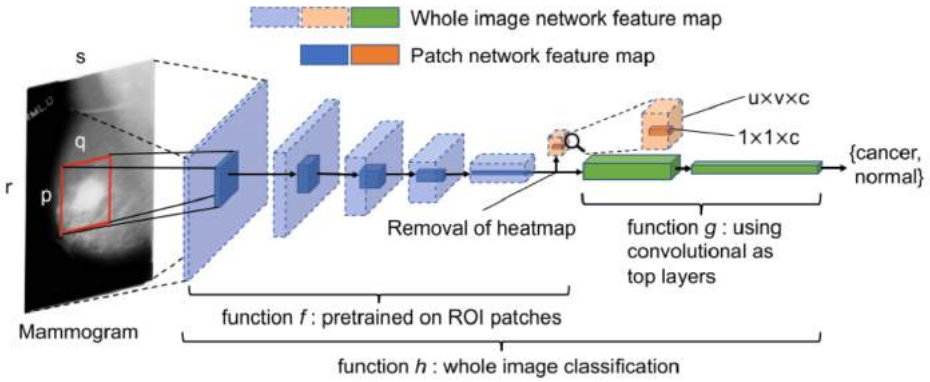


Figure B.3: The architecture of the end2end model (L. Shen et al., 2019)

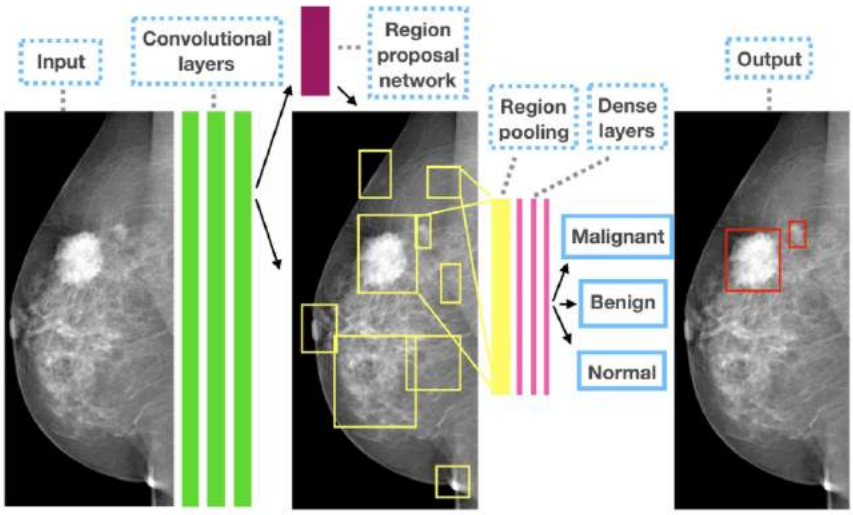


Figure B.4: The architecture of the faster rcnn model (Ribli et al., 2018)

References

Abdelhafiz, D., Yang, C., Ammar, R., & Nabavi, S. (2019). Deep convolutional neural networks for mammography: Advances, challenges and applications. *BMC Bioinformatics*, *20*(Suppl 11). https://doi.org/10.1186/s12859-019-2823-4

Alamin, I. M. J., Jeberson, W., & Bajaj, H. (2016). Improved Region Growing based Breast Cancer Image Segmentation. *International Journal of Computer Applications*, *135*(February). https://doi.org/10.5120/ijca2016908244

Al-masni, M. A., Al-antari, M., Park, J., Gi, G., Kim, T., Rivera, P., Valarezo, E., Han, S. .-., & Kim, T. .-. (2017). Detection and classification of the breast abnormalities in digital mammograms via regional Convolutional Neural Network. *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*.

Bakker, M. F., de Lange, S. V., Pijnappel, R. M., Mann, R. M., Peeters, P. H., Monninkhof, E. M., Emaus, M. J., Loo, C. E., Bisschops, R. H., Lobbes, M. B., de Jong, M. D., Duvivier, K. M., Veltman, J., Karssemeijer, N., de Koning, H. J., van Diest, P. J., Mali, W. P., van den Bosch, M. A., Veldhuis, W. B., & van Gils, C. H. (2019). Supplemental MRI Screening for Women with Extremely Dense Breast Tissue. *New England Journal of Medicine*, *381*(22), 2091–2102. https://doi.org/10.1056/nejmoa1903986

Berber, T., Alpkocak, A., Balci, P., & Dicle, O. (2013). Breast mass contour segmentation algorithm in digital mammograms. *Computer Methods and Programs in Biomedicine*, *110*(2), 150–159. https://doi.org/10.1016/j.cmpb.2012.11.003

Boyd, N. F., Guo, H., Martin, L. J., Sun, L., Stone, J., Fishell, E., Jong, R. A., Hislop, G., Chiarelli, A., Minkin, S., & Yaffe, M. J. (2007). Mammographic Density and the Risk and Detection of Breast Cancer. *Breast Diseases*, *356*(3), 227–236. https://doi.org/10.1016/S1043-321X(07)80400-0

Brem, R. F., Ioffe, M., Rapelyea, J. A., Yost, K. G., Weigert, J. M., Bertrand, M. L., & Stern, L. H. (2009). Invasive lobular carcinoma: Detection with mammography, sonography, MRI, and breast-specific gamma imaging. *American Journal of Roentgenology*, *192*(2), 379–383. https://doi.org/10.2214/AJR.07.3827

Carneiro, G., Nascimento, J., & Bradley, A. P. (2017). Automated Analysis of Unregistered Multi-View Mammograms with Deep Learning. *IEEE Transactions on Medical Imaging*, *36*(11), 2355–2365. https://doi.org/10.1109/TMI.2017.2751523

Chen, Z., Yang, J., Li, S., Lv, M., Shen, Y., Wang, B., Li, P., Yi, M., Zhao, X., Zhang, L., Wang, L., & Yang, J. (2017). Invasive lobular carcinoma of the breast: A special histological type compared with invasive ductal carcinoma. *PLoS ONE*, *12*(9), 1–17. https://doi.org/10.1371/journal.pone.0182397

Chougrad, H., Zouaki, H., & Alheyane, O. (2018). Deep Convolutional Neural Networks for breast cancer screening. *Computer Methods and Programs in Biomedicine*, *157*(January), 19–30. https://doi.org/10.1016/j.cmpb.2018.01.011

Christgen, M., Steinemann, D., Kühnle, E., Länger, F., Gluz, O., Harbeck, N., & Kreipe, H. (2016). Lobular breast cancer: Clinical, molecular and morphological characteristics. *Pathology Research and Practice*, *212*(7), 583–597. https://doi.org/10.1016/j.prp.2016.05.002

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves : A Nonparametric Approach Author (s): Elizabeth R . DeLong , David M . DeLong and Daniel L . Clarke-Pearson Published by : International Biometric Society Stable. *Biometrics*, *44*, 837–845.

Dembrower, K., Lindholm, P., & Strand, F. (2020). A Multi-million Mammography Image Dataset and Population-Based Screening Cohort for the Training and Evaluation of Deep Neural Networks—the Cohort of Screen-Aged Women (CSAW). *Journal of Digital Imaging*, *33*, 408–413. https://doi.org/10.1007/s10278-019-00278-0

Desmedt, C., Zoppoli, G., Sotiriou, C., & Salgado, R. (2017). Transcriptomic and genomic features of invasive lobular breast cancer. *Seminars in Cancer Biology*, *44*, 98–105. https://doi.org/10.1016/j.semcancer.2017.03.007

Dhungel, N., Carneiro, G., & Bradley, A. P. (2016). The automated learning of deep features for breast mass classification from mammograms. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 106–114. https://doi.org/10.1007/978-3-319-46723-8_13

Dogra, A., Goyal, B., & K, K. (2019). A Brief Review of Breast Cancer Detection via Computer Aided Deep Learning Methods. *International Journal of Engineering Research & Technology*, *8*(12), 326–331. https://doi.org/10.17577/ijertv8is120191

Duggento, A., Aiello, M., Cavaliere, C., Cascella, G. L., Cascella, D., Conte, G., Guerrisi, M., & Toschi, N. (2019). An Ad Hoc Random Initialization Deep Neural Network Architecture for Discriminating Malignant Breast Cancer Lesions in Mammographic IMages. *Contrast Media & Molecular Imaging*, *2019*. https://doi.org/10.1109/EMBC.2019.8856740

Ferrari, R. J., Frère, A. F., Rangayyan, R. M., Desautels, J. E., & Borges, R. A. (2004). Identification of the breast boundary in mammograms using active contour models. *Medical and Biological Engineering and Computing*, *42*(2), 201–208. https://doi.org/10.1007/BF02344632

Geras, K. J., Wolfson, S., Shen, Y., Wu, N., Gene Kim, S., Kim, E., Heacock, L., Parikh, U., Moy, L., & Cho, K. (2017). High-resolution breast cancer screening with multi-view deep convolutional neural networks. *arXiv*, 1–9.

Gonzalez, R. C., & Woods, R. E. (2009). *Digital Image Processing - Third Edition*. Pearson Education.

Grgic, M., Delac, K., & Ghanbari, M. (2009). A Survey of Image Processing Algorithms in Digital Mammography. *Recent advances in multimedia signal processing and communications* (pp. 631–657).

Guiu, S., Wolfer, A., Jacot, W., Fumoleau, P., Romieu, G., Bonnetain, F., & Fiche, M. (2014). Invasive lobular breast cancer and its variants: How special are they for systemic therapy decisions? *Critical Reviews in Oncology/Hematology*, *92*(3), 235–257. https://doi.org/10.1016/j.critrevonc.2014.07.003

Hemalatha, R., Thamizhvani, T., Josephin Arockia Dhivya, A., Joseph, J. E., Babu, B., & Chandrasekaran, R. (2018). Active Contour Based Segmentation Techniques for Medical Image Analysis. *Medical and biological image analysis* (pp. 137–144). http://www.intechopen.com/books/trends-in-telecommunications-technologies / gps - total - electron - content - tec - %20prediction - at - ionosphere - layer - over - the - equatorial - region%7B%5C%%7D0AInTec%7B%5C%%7D0Ahttp://www.asociatiamhc.ro/wp-content/uploads/2013/11/Guide-to-Hydropower.pdf

Hilleren, D. J., Andersson, I. T., Lindholm, K., & Linell, F. S. (1991). Breast Lobular Carcinoma : Mammographic Findings in a 10-year Experience. *Radiology*, *178*(1), 149–154.

Hwang, S., & Kim, H. E. (2016). Self-Transfer Learning for Fully Weakly Supervised Object Localization. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 9901 LNCS*, 239–246. https://doi.org/10.1007/978-3-319-46723-8_28

IKNL. (n.d.). Nederlandse kankerregistratie (NKR). Retrieved November 30, 2021, from iknl.nl/kankersoorten/borstkanker/registratie/incidentie

Jiménez-Gaona, Y., Rodríguez-Álvarez, M. J., & Lakshminarayanan, V. (2020). Deep-Learning-Based Computer-Aided Systems for Breast Cancer Imaging : A Critical Review. *Applied Sciences*, *10*(8298).

John Hopkins University. (2021). Types of Breast Cancer. Retrieved November 30, 2021, from pathology.jhu.edu/breast/types-of-breast-cancer

Johnson, K., Sarma, D., & Hwang, E. S. (2015). Lobular breast cancer series: Imaging. *Breast Cancer Research*, *17*(1), 1–8. https://doi.org/10.1186/s13058-015-0605-0

Kavanagh, A. M., Giles, G. G., Mitchell, H., & Cawson, J. N. (2000). The sensitivity, specificity, and positive predictive value of screening mammography and symptomatic status. *Journal of Medical Screening*, *7*(2), 105–110. https://doi.org/10.1136/jms.7.2.105

Kerlikowske, K., Grady, D., Barclay, J., Sickles, E. A., & Ernster, V. (1996). Effect of age, breast density, and family history on the sensitivity of first screening mammography. *Journal of the American Medical Association*, *276*(1), 33–38. https://doi.org/10.1001/jama.276.1.33

Kingma, D. P., & Ba, J. L. (2015). Adam: A Method for Stochastic Optimization. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 1–15.

Kom, G., Tiedeu, A., & Kom, M. (2007). Automated detection of masses in mammograms by local adaptive thresholding. *Computers in Biology and Medicine*, *37*(1), 37–48. https://doi.org/10.1016/j.compbiomed.2005.12.004

Kooi, T., Litjens, G., van Ginneken, B., Gubern-Mérida, A., Sánchez, C. I., Mann, R., den Heeten, A., & Karssemeijer, N. (2017). Large Scale Deep Learning for Computer Aided Detection of Mammographic Lesions. *Medical Image Analysis*, *35*, 303–312. https://doi.org/10.1016/j.media.2016.07.007

Lehman, C. D., Arao, R. F., Sprague, B. L., Lee, J. M., Buist, D. S. M., Kerlikowske, K., Henderson, L. M., Tosteson, A. N. A., Rauscher, G. H., & Miglioretti, D. L. (2017). National Performance Benchmarks for Modern Screening Digital Mammography. *Radiology*, *283*(1), 49–58.

Lehman, C. D., Wellman, R. D., Buist, D. S. M., Kerlikowske, K., Tosteson, A. N. A., & Miglioretti, D. L. (2015). Diagnostic Accuracy of Digital Screening Mammography With and Without Computer-Aided Detection. *JAMA Intern Med.*, *175*(11), 1828–1837.

Liu, J., Liu, X., Chen, J., & Tang, J. (2011). Mass Segmentation in Mammograms Based on Improved Level Set and Watershed Algorithm. *Advanced Intelligent Computing Theories and Applications*, 502–508.

Liu, K., Shen, Y., Wu, N., Chłdowski, J., Fernandez-Granda, C., & Geras, K. J. (2021). Weakly-supervised High-resolution Segmentation of Mammography Images for Breast Cancer Diagnosis. *Proceedings of Machine Learning Research*, *1*(22). http://arxiv.org/abs/2106.07049

Masud, R., Al-Rei, M., & Lokker, C. (2019). Computer-Aided Detection for Breast Cancer Screening in Clinical Settings : Scoping Review. *JMIR Medical Informatics*, *7*(3). https://doi.org/10.2196/12660

Mckinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G. C., Darzi, A., Etemadi, M., Garcia-vicente, F., Gilbert, F. J., Halling-brown, M., Hassabis, D., Jansen, S., & Karthikesalingam, A. (2020). International evaluation of an AI system for breast cancer screening. *Nature*, *577*(January 2020), 89–94. https://doi.org/10.1038/s41586-019-1799-6

Mello-thoms, C. (2020). The Path to Implementation of Artificial Intelligence in Screening Mammography Is Not All That Clear. *JAMA Network Open*, *3*(3), 9–11. https://doi.org/10.2214/AJR.19.21346

Perry, N., Broeders, M., de Wolf, C., Törnberg, S., Holland, R., & von Karsa, L. (2008). European guidelines for quality assurance in breast cancer screening and diagnosis. *Oncology in Clinical Practice*, *4*(2), 74–86.

Pisano, E. D., Cole, E. B., Hemminger, B. M., Yaffe, M. J., Aylward, S. R., Maidment, A. D. A., Johnston, R. E., Williams, M. B., Niklason, L. T., Conant, E. F., Fajardo, L. L., Kopans, D. B., Brown, M. E., & Pizer, S. M. (2000). Algorithms for Digital Mammography : A Pictorial Essay. *RadioGraphics*, *20*(5), 1479–1491.

Preim, B., & Botha, C. (2014). Image Analysis for Medical Visualization. *Visual computing for medicine* (Second Edi, pp. 111–175). Morgan Kaufmann - Elsevier. https://doi.org/10.1016/c2011-0-05785-x

Ribli, D., Horváth, A., Unger, Z., Pollner, P., & Csabai, I. (2018). Detecting and classifying lesions in mammograms with Deep Learning. *Scientific Reports*, *8*(4165), 1–7. https://doi.org/10.1038/s41598-018-22437-z

Savelli, B., Bria, A., Molinara, M., Marrocco, C., & Tortorella, F. (2019). A multi-context CNN ensemble for small lesion detection. *Artificial Intelligence in Medicine*, *103*(101749). https://doi.org/10.1016/j.artmed.2019.101749

Sechopoulos, I., & Mann, R. M. (2020). Stand-alone artificial intelligence - The future of breast cancer screening? *The Breast*, *49*, 254–260. https://doi.org/10.1016/j.breast.2019.12.014

Shen, L., Margolies, L. R., Rothstein, J. H., Fluder, E., McBride, R., & Sieh, W. (2019). Deep Learning to Improve Breast Cancer Detection on Screening Mammography. *Scientific Reports*, *9*(12495). https://doi.org/10.1038/s41598-019-48995-4

Shen, R., Yan, K., Tian, K., Jiang, C., & Zhou, K. (2019). Breast mass detection from the digitized X-ray mammograms based on the combination of deep active learning and self-paced learning. *Future Generation Computer Systems*, *101*, 668–679. https://doi.org/10.1016/j.future.2019.07.013

Shen, Y., Wu, N., Phang, J., Park, J., Liu, K., Tyagi, S., Heacock, L., Kim, S. G., Moy, L., Cho, K., & Geras, K. J. (2019). An interpretable classifier for high-resolution breast cancer screening images utilizing weakly supervised localization. *Machine Learning in Medical Imaging: 10th international Workshop, MLMI 2019, October 13*, 18. https://doi.org/10.1016/j.media.2020.101908

Shi, J., Sahiner, B., Chan, H.-P., Ge, J., Hadjiiski, L., Helvie, M. A., Nees, A., Wu, Y. T., Wei, J., Zhou, C., Zhang, Y., & Cui, J. (2008). Characterization of mammographic masses based on level set segmentation with new image features and patient information. *Medical Physics*, *35*(1), 280–290. https://doi.org/10.1118/1.2820630

Singh, V. K., Rashwan, H. A., Romani, S., Akram, F., Pandey, N., Mostafa Kamal Sarker, M., Saleh, A., Arenas, M., Arquez, M., Puig, D., & Torrents-Barrena, J. (2019). Breast tumor segmentation and shape classification in mammograms using generative adversarial and convolutional neural network. *Expert Systems with Applications*, *139*(112855).

Strickland, R. N., & Hahn, H. I. (1996). Wavelet Transforms for Detecting Microcalcifications in Mammograms. *Proceedings - International Conference on Image Processing, ICIP*, *15*(2), 218–229. https://doi.org/10.1109/ICIP.1994.413344

Suckling, J., Parker, J., Dance, D., Astley, S., Hutt, I., Boggis, C., & Ricketts, I. (2015). Mammographic Image Analysis Society (MIAS) database v1.21 [Dataset].

Sun, W., Tseng, T.-L. B., Zhang, J., & Qian, W. (2016). Enhancing deep convolutional neural network scheme for breast cancer diagnosis with unlabeled data. *Computerized Medical Imaging and Graphics*, *57*, 4–9. https://doi.org/10.1016/j.compmedimag.2016.07.004

von Euler-Chelpin, M., Lillholm, M., Napolitano, G., Vejborg, I., Nielsen, M., & Lynge, E. (2018). Screening mammography: benefit of double reading by breast density. *Breast Cancer Research and Treatment*, *171*(3), 767–776. https://doi.org/10.1007/s10549-018-4864-1

Wadhwa, A., Sullivan, J. R., & Gonyo, M. B. (2016). Missed Breast Cancer: What Can We Learn? *Current Problems in Diagnostic Radiology*, *45*(6), 402–419. https://doi.org/10.1067/j.cpradiol.2016.03.001

Wanders, J. O., Holland, K., Veldhuis, W. B., Mann, R. M., Pijnappel, R. M., Peeters, P. H., van Gils, C. H., & Karssemeijer, N. (2017). Volumetric breast density affects performance of digital screening mammography. *Breast Cancer Research and Treatment*, *162*(1), 95–103. https://doi.org/10.1007/s10549-016-4090-7

Weiss, K., Khoshgoftaar, T. M., & Wang, D. (2016). A survey of transfer learning. *Journal of Big Data*, *3*(9). https://doi.org/10.1186/s40537-016-0043-6

Wong, D. J., Gandomkar, Z., Wu, W.-J., Zhang, G., Gao, W., He, X., Wang, Y., & Reed, W. (2020). Artificial intelligence and convolution neural networks assessing mammographic images: a narrative literature review. *Journal of Medical Radiation Sciences*, *67*(2), 134–142. https://doi.org/10.1002/jmrs.385

Wu, N., Geras, K. J., Shen, Y., Su, J., Kim, S. G., Kim, E., Wolfson, S., Moy, L., & Cho, K. (2018). Breast density classification with deep convolutional neural networks. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, *April*, 6682–6686. https://doi.org/10.1109/ICASSP.2018.8462671

Wu, N., Phang, J., Park, J., Shen, Y., Huang, Z., Zorin, M., Jastrzebski, S., Fevry, T., Katsnelson, J., Kim, E., Wolfson, S., Parikh, U., Gaddam, S., Lin, L. L. Y., Ho, K., Weinstein, J. D., Reig, B., Gao, Y., Toth, H., … Geras, K. J. (2020). Deep Neural Networks Improve Radiologists' Performance in Breast Cancer Screening. *IEEE Transactions on Medical Imaging*, *39*(4), 1184–1194. https://doi.org/10.1109/TMI.2019.2945514

Wu, N., Phang, J., Park, J., Shen, Y., Kim, S. G., Heacock, L., Moy, L., Cho, K., & Geras, K. J. (2019). The NYU Breast Cancer Screening Dataset v1.0. https://cs.nyu.edu/%7B~%7Dkgeras/reports/datav1.0.pdf

Young, I. T., Gerbrands, J. J., & van Vliet, L. J. (2007). Fundamentals of Image Processing. https://doi.org/10.1002/9783527635245.ch4

Zhang, Y., Tomuro, N., Furst, J., & Stan Raicu, D. (2010). Image enhancement and edge-based mass segmentation in mammogram. *Proceedings of SPIE - The International Society for Optical ENgineering*, *7623*(March 2010). https://doi.org/10.1117/12.844492