

COMPUTER-AIDED DIAGNOSIS
IN SCREENING MRI OF WOMEN
WITH
EXTREMELY DENSE BREASTS



COMPUTER-AIDED DIAGNOSIS
IN SCREENING MRI OF WOMEN
WITH
EXTREMELY DENSE BREASTS

Erik Verburg

Cover image: © Hans Jochem Bakker

Printing: Ipskamp Printing

ISBN: 978-94-6421-899-2

DOI: 10.33540/1429

URL: <https://doi.org/10.33540/1429>

© 2022 E. Verburg, The Netherlands. All rights reserved. No parts of this thesis may be reproduced, stored in a retrieval system or transmitted in any form or by any means without permission of the author.

Alle rechten voorbehouden. Niets uit deze uitgave mag worden vermenigvuldigd, in enige vorm of op enige wijze, zonder voorafgaande schriftelijke toestemming van de auteur

COMPUTER-AIDED DIAGNOSIS IN SCREENING MRI OF WOMEN WITH EXTREMELY DENSE BREASTS

Computer-ondersteuning bij de diagnose van MRI beelden uit de screening van
vrouwen met zeer dicht borstweefsel
(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de
Universiteit Utrecht
op gezag van de
rector magnificus, prof.dr. H.R.B.M. Kummeling,
ingevolge het besluit van het college voor promoties
in het openbaar te verdedigen op

dinsdag 22 november 2022 des middags te 4.15 uur

door

Erik Verburg
geboren op 1 maart 1986
te Zwolle

Promotor:

Prof.dr.ir. M.A. Viergever

Copromotor:

Dr. K.G.A. Gilhuijs

Beoordelingscommissie:

Prof.dr. D.L. Oberski

Prof.dr. P.J. van Diest (voorzitter)

Prof.dr. C.T.W. Moonen

Prof.dr. P.A.N. Bosman

Dr. M.R.Moman

Table of contents

1	Introduction	7
2	Knowledge-based and deep learning-based automated chest wall segmentation in Magnetic Resonance Images of extremely dense breasts	17
3	Deep Learning for Automated Triaging of 4 581 breast MRIs from the DENSE Trial	47
4	Computer-aided diagnosis in multi-parametric MRI screening of women with extremely dense breasts to reduce false positive diagnoses	71
5	Validation of combined Deep-learning Triaging and Computer-aided Diagnosis in 2,901 Breast MRI Examinations from the Second Screening Round of the DENSE Trial	97
6	Discussion	119
7	Appendix	131
	English summary	132
	Nederlandse samenvatting (Dutch summary)	136
	Acknowledgments	140
	Dankwoord	141
	List of publications	144
	Curriculum Vitae	145

Chapter 1

Introduction



1.1 Breast density and the DENSE trial

Breasts contain glandular, connective, and fat tissue. Breast density is a term that describes the relative amount of these different types of breast tissue as seen on a mammogram. According to the American College of Radiology there are four types of breast density, almost entirely fat (A), scattered areas of fibroglandular density (B), heterogeneously dense (C) and extremely dense (D) (Figure 1.1). Breast density can be measured visually by the radiologist[1], or semi- or fully automated using software[2–4].

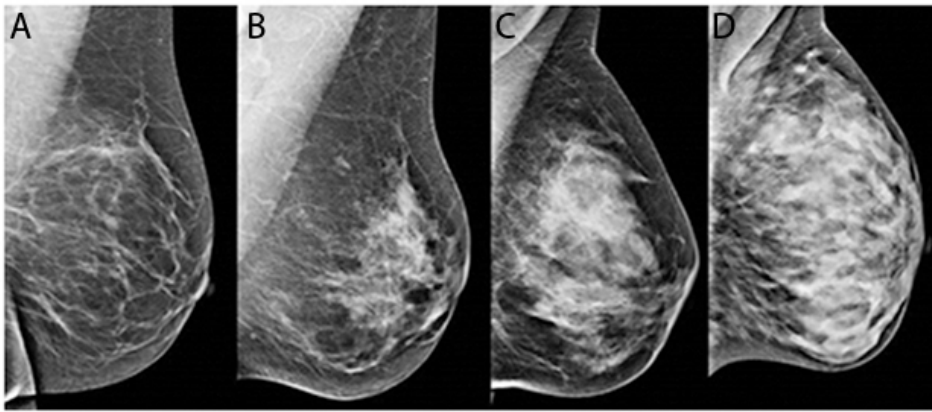


Figure 1.1: examples of mammography of the four types of breast density, almost entirely fat (A), scattered areas of fibroglandular density (B), heterogeneously dense (C) and extremely dense (D). (source: UMC Utrecht)

Extremely dense breasts and heterogeneously dense breasts have relatively high amounts of glandular tissue and fibrous connective tissue and relatively low amounts of fatty breast tissue. Approximately 29% of the Dutch women between 50 and 75 years have heterogeneously dense breasts and 8% have extremely dense breasts[5]. Compared to entirely fatty breasts, women with extremely dense breasts have a 2-6 times higher risk of developing breast cancer[6–10]. Moreover, it is difficult to detect tumors in these breasts using mammography due to the low contrast between the fibroglandular tissue and tumor tissue[5, 11, 12]. Consequently, more sensitive techniques such as dynamic contrast-enhanced magnetic resonance imaging (DCE MRI) are investigated to screen these women[13, 6]. However, DCE MRI is associated with varying specificity to discriminate between malignant (cancer) and benign (no cancer) lesions[13, 14], while it produces more

imaging data than a mammographic examination. The Dense Tissue and Early Breast Neoplasm Screening (DENSE) trial investigated whether MRI in addition to the current screening practice (mammography) results in less interval carcinomas. For this purpose, DENSE has two arms in a prospective randomized controlled trial with three screening rounds. In short, if the screening mammogram shows extremely dense breasts, women were randomly assigned to the control arm (i.e., conventional screening mammography only) or the MRI arm where they receive additional bi-annual MRI screening upon invitation. The invitation also informed women of their breast density and the associated risk of developing breast cancer. In the first screening round 8061 women were invited for additional MRI screening, 5276 (66%) were interested and 4783 (59.3%) completed the MRI examination[15]. All MRI screening images were assessed by experienced breast MR radiologists. In total, 4329 participants had normal breasts on MRI, the remaining 454 (9.5%) participants were invited for a repeat MRI after six months or biopsy. Cancer was found in 79 (17.5%) of 454 referred women (i.e., 16.5/1000 women). No cancer (i.e. false-positive referral) occurred in 375 (82.5%) women[16]. The results of the first screening round showed that additional MRI for women with extremely dense breasts significantly reduced the number of interval cancers compared to the control arm[16]. To detect 79 women with breast cancer, 375 healthy women received an invitation for further work-up. Because approximately 225.000 women may be eligible for additional MRI screening due to dense breasts in the Netherlands alone[17], the number of women with false-positive findings may increase more than twenty-fold.

1.2 MRI screening

The MR images in DENSE were obtained using a multiparametric MRI protocol, comprised of T1-weighted dynamic contrast-enhanced image series (DCE), diffusion weighted images (DWI) and a T2-weighted sequence[6]. The DCE sequence is a series of images during which contrast agent is injected. Contrast dynamics in the breast tissue are visualized by acquiring one high spatial resolution three dimensional (3D) image before contrast injection, multiple (15-20) 3D images with a high temporal resolution and lower spatial resolution during contrast uptake, followed by 4 to 5 3D images with high spatial resolution after the contrast is injected. In general, contrast uptake in lesions is higher compared to that in the surrounding normal tissue. Hence, the DCE images series with high

spatial resolution give information about the shape, volume and heterogeneity of breast lesions[13]. The dynamics of contrast uptake contain important information to characterize breast lesions. Three types of contrast dynamics are defined (Figure 1.2). Type 1 curves show a persistent uptake of contrast agent into the lesion, type 2 curves show a plateau of contrast uptake after circa 90 s and type 3 show washout of the contrast agent after 90 s. A type 1 curve is an indication for a benign lesion, type 2 curve lesions have an intermediate risk of malignancy and lesions showing contrast curve type 3 are probably malignant (Figure 1.3)

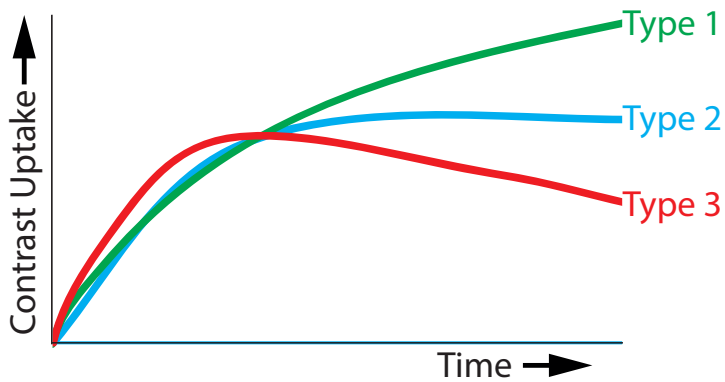


Figure 1.2: Three typical curves of contrast uptake over time in breast lesions. Lesions showing type 1 are probably benign, type 2 lesions have intermediate risk and type 3 lesions are probably malignant.

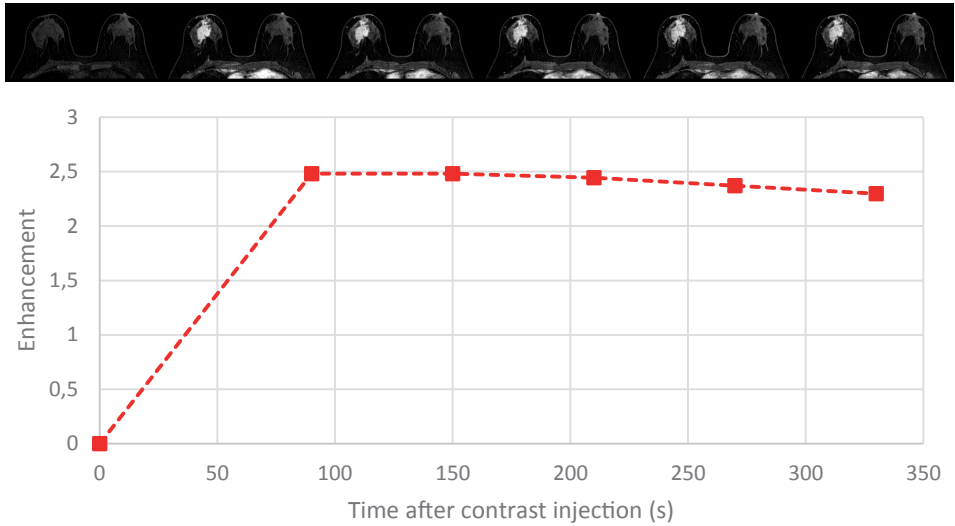


Figure 1.3: example of a DCE image series of a malignant lesion showing type 3 enhancement. The image most left shows the T1-weighted MRI without contrast. Ninety seconds after contrast injection the second image is made, showing the fast contrast uptake in the lesion in the right breast (left in image). In the next four images acquired 150 s, 210 s, 270 s, and 330 s after contrast injection, the image intensity at the location of the lesion reduces. The graph shows the median image intensity in the region of the malignant lesion at each time point.

DWI shows the diffusion of water in the breast tissue. Using these images it is possible to calculate an apparent diffusion coefficient (ADC) which is a measure for the movement of water inside the tissue. In general, less diffusion, movement of water, is present in malignant tissue due to the fast irregular growth of cancer tissue (Figure 1.4).

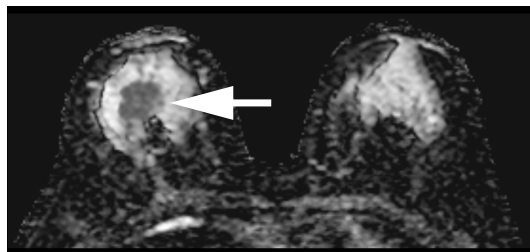


Figure 1.4: ADC map created of DWI images. In the dark gray parts, diffusion is restricted (arrow). In the bright parts, water diffusion is less restricted. In this example, the malignant lesion (arrow) shows less diffusion than the surrounding normal tissue.

T1-weighted images sometimes are referred to as “anatomical scan” because different tissues types are well distinguishable. T2-weighted images show the

presence of water in the tissue, and for example can be used to detect cysts, which are filled with fluid mostly consisting of water[13].

Machine learning and Computer Aided Diagnosis

This thesis focuses on new techniques to automatically reduce false-positive findings in breast MRI screening. Machine learning is a technique for recognizing patterns in, e.g., medical images. The calculation power of current computers allows us to use complex machine learning methods to be used in, for example, computer-aided diagnosis. In computer aided diagnosis (CAD) the outcome of the computer supports the professional in decisions making (Figure 1.5).

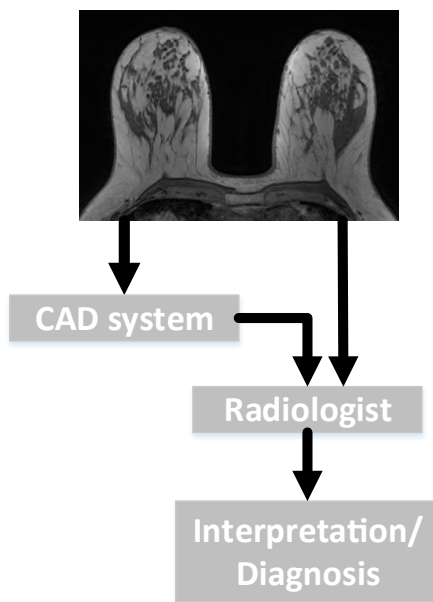


Figure 1.5: Schematic diagram of a CAD system for medical image interpretation

Computers are able to calculate multiple properties of images, called features, and use these features to predict whether, for example, a breast lesion is benign or malignant. These features can be properties describing shape, volume, color, intensity, intensity changes over time and so on. Machine learning can be used for multiple purposes. In medical image assessment it is mostly used for detection, quantification or classification. For example, it is used to detect whether a lesion is present or to delineate different tissue types (segmentation). Classification can

be used to predict whether a lesion is benign or malignant. In this thesis we develop new methods to apply on images acquired using a multi-parametric MRI sequence.

1.3 Outline of this thesis

In this thesis we will develop and validate artificial intelligence for breast MRI in women with extremely dense breasts to assist in the identification and characterization of breast lesions, thus pursuing reduction of false positive referrals. One of the first steps in image analysis is the definition of the region of interest. In MR images, breast tissue is surrounded by skin and the chest wall, consisting of the pectoral muscles and chest bones. The border between air and breast tissue is easily distinguishable and detectable using simple techniques. The detection of the chest wall is, however, more challenging. Multiple methods showed difficulties to delineate the chest wall, especially in MRI of extremely dense breasts, due to the lack of contrast between glandular tissue and muscle tissue.

In **Chapter 2** we develop and compare multiple automated methods, dedicated to segment the chest wall in breast MRI of extremely dense breasts.

In **Chapter 3** we train and validate a computer aided triaging (CAT) method based on deep learning, which dismisses breast MRI examinations without lesions from further radiological review, without dismissing examinations in which malignant lesions were present.

In **Chapter 4**, we develop CAD to reduce the number of false positive referrals without missing any malignant disease.

In **Chapter 5**, the CAT method of **Chapter 3** and the CAD method of **Chapter 4** are combined and applied on the MRI data acquired during the second screening round of the dense trial to investigate the performance of the combination of the methods.

In **Chapter 6**, we discuss the benefits of the developed computer aided methods, the opportunities of these methods for MRI screening and the challenges for the introduction of the methods in the clinical workflow.

References

- [1] E. Morris et al., "Acr bi-rads® atlas, breast imaging reporting and data system," *Reston, VA: American College of Radiology*, pp. 56–71, 2013.
- [2] S. Ciatto et al., "A first evaluation of breast radiological density assessment by quantra software as compared to visual classification," *The Breast*, vol. 21, no. 4, pp. 503–506, 2012.
- [3] S. van Engeland et al., "Volumetric breast density estimation from full-field digital mammograms," *IEEE Trans Med Imaging*, vol. 25, no. 3, pp. 273–82, 2006.
- [4] C. D. Lehman et al., "Mammographic breast density assessment using deep learning: Clinical implementation," *Radiology*, vol. 290, no. 1, pp. 52–58, 2019.
- [5] J. O. Wanders et al., "Volumetric breast density affects performance of digital screening mammography," *Breast Cancer Res Treat*, vol. 162, no. 1, pp. 95–103, 2017.
- [6] M. J. Emaus et al., "Mr imaging as an additional screening modality for the detection of breast cancer in women aged 50-75 years with extremely dense breasts: The dense trial study design," *Radiology*, vol. 277, no. 2, pp. 527–37, 2015.
- [7] E. R. Price et al., "The california breast density information group: a collaborative response to the issues of breast density, breast cancer risk, and breast density notification legislation," *Radiology*, vol. 269, no. 3, pp. 887–92, 2013.
- [8] V. A. McCormack and I. dos Santos Silva, "Breast density and parenchymal patterns as markers of breast cancer risk: a meta-analysis," *Cancer Epidemiol Biomarkers Prev*, vol. 15, no. 6, pp. 1159–69, 2006.
- [9] N. F. Boyd et al., "Mammographic density and the risk and detection of breast cancer," *N Engl J Med*, vol. 356, no. 3, pp. 227–36, 2007.
- [10] C. M. Vachon et al., "Mammographic density, breast cancer risk and risk prediction," *Breast cancer research : BCR*, vol. 9, no. 6, pp. 217–217, 2007.
- [11] P. A. Carney, D. L. Miglioretti, B. C. Yankaskas, and et al., "Individual and combined effects of age, breast density, and hormone replacement therapy use on the accuracy of screening mammography," *Annals of Internal Medicine*, vol. 138, no. 3, pp. 168–175, 2003.
- [12] K. Kerlikowske, "The mammogram that cried wolfe," *New England Journal of Medicine*, vol. 356, no. 3, pp. 297–300, 2007.
- [13] R. M. Mann, C. K. Kuhl, K. Kinkel, and C. Boetes, "Breast mri: guidelines from the european society of breast imaging," *European radiology*, vol. 18, no. 7, pp. 1307–1318, 2008.
- [14] E. Warner et al., "Systematic review: Using magnetic resonance imaging to screen women at high risk for breast cancer," *Annals of Internal Medicine*, vol. 148, no. 9, pp. 671–679, 2008.
- [15] S. V. de Lange et al., "Reasons for (non)participation in supplemental population-based mri breast screening for women with extremely dense breasts," *Clinical Radiology*, vol. 73, no. 8, pp. 759.e1–759.e9, 2018.

[16] M. F. Bakker et al., "Supplemental mri screening for women with extremely dense breast tissue," *New England Journal of Medicine*, vol. 381, no. 22, pp. 2091–2102, 2019.

[17] IKNL, "Monitor bevolkingsonderzoek borstkanker," 2018.

Chapter 2

Knowledge-based and deep learning-based automated chest wall segmentation in Magnetic Resonance Images of extremely dense breasts

Based on: **Erik Verburg**, Jelmer M. Wolterink, Stephanie N. de Waard, Ivana Išgum, Carla H. van Gils, Wouter B. Veldhuis, Kenneth G. A. Gilhuijs; Knowledge-based and deep learning-based automated chest wall segmentation in Magnetic Resonance Images of extremely dense breasts, *Medical Physics*, 2019, Volume 46 (4405-4416)

2.1 Abstract

Purpose

Segmentation of the chest wall, is an important component of methods for automated analysis of breast MRI. Methods reported to date show promising results but have difficulties delineating the muscle border correctly in breasts with a large proportion of fibroglandular tissue (i.e., dense breasts). Knowledge-based methods as well as methods based on deep learning have been proposed, but a systematic comparison of these approaches within one cohort of images is currently lacking. Therefore, we developed a knowledge-based method and a deep learning method for segmentation of the chest wall in magnetic resonance imaging (MRI) of dense breasts and compared their performances.

Methods

Two automated methods were developed, an optimized knowledge-based method (KBM) incorporating heuristics aimed at shape, location and gradient features, and a deep learning-based method (DLM) using a dilated convolution neural network. A dataset of 115 T1-weighted MR images was randomly selected from MR images of women with extremely dense breasts (ACR BI-RADS category 4) participating in a screening trial of women (mean age 56.6 years, range 49.5-75.2 years) with dense breasts. Manual segmentations of the chest wall, acquired under supervision of an experienced breast radiologist, were available for all datasets. Both methods were optimized using the same randomly selected 36 MRI datasets from a total of 115 datasets. Each MR dataset consisted of 179 transversal images with voxel size $0.64 \times 0.64 \times 1.00 \text{ mm}^3$. In the remaining 79 datasets, the results of both segmentation methods were qualitatively evaluated. A radiologist reviewed segmentation results of both methods in all transversal images ($n=141$) and determined whether the result would impact the ability to accurately determine volume of fibroglandular and fatty tissue and whether segmentations masked breast regions that might harbor lesions. When no relevant deviation was detected, the result was considered successful. In addition, all segmentations were quantitatively assessed using the Dice similarity coefficient (DSC) and Hausdorff distance (HD), 95th percentile of the Hausdorff distance (HD95), false positive fraction (FPF) and false negative fraction (FNF) metrics.

Results

According to the radiologist's evaluation, the DLM had a significantly higher success rate than the KBM (81.6% vs. 78.4%, $p < 0.01$). The success rate was further improved to 92.1% by combining both methods. Similarly, the DLM had significantly lower values for FNF (0.003 ± 0.003 vs. 0.009 ± 0.011 , $p < 0.01$) and HD95 (2.58 ± 1.78 mm vs. 3.37 ± 2.11 , $p < 0.01$). However, the KBM resulted in a significantly lower FPF than the DLM (0.018 ± 0.009 vs. 0.030 ± 0.009 , $p < 0.01$). There was no significant difference between the KBM and DLM in terms of DSC (0.982 ± 0.006 vs. 0.984 ± 0.008 , $p = 0.08$) or HD (24.14 ± 20.69 mm vs. 12.81 ± 27.28 mm, $p = 0.05$).

Conclusion

Both optimized knowledge-based and deep learning-based method showed good results to segment the pectoral muscle in women with dense breasts. Qualitatively assessed, the DLM was the most robust method. A quantitative comparison, however, did not indicate a preference for one method over the other.

2.2 Introduction

Breast cancer is the most common type of cancer in women in western countries. Detecting breast cancers at an early stage yields a survival benefit [1]. Mammography screening programs exist in many countries. The sensitivity of mammography is lower, however, in women with dense breasts (i.e., American College of Radiology (ACR) class 3 and ACR class 4), which comprises approximately 40% of the population[2–4]. Moreover, the risk of developing breast cancer is two- to six-fold higher in women who have a large proportion of fibroglandular tissue in their breasts (i.e., dense breasts on mammography)[2, 5, 6]. Consequently, more sensitive techniques such as dynamic contrast-enhanced magnetic resonance imaging (DCE MRI) are investigated for screening these women[2]. However, DCE MRI is associated with varying specificity to discriminate between benign and malignant lesions[7], while it produces more imaging data than a mammographic examination. Hence, computer-aided diagnosis (CAD) of DCE MRI is becoming increasingly important to reduce workload and biopsies on benign lesion.

A typical first step in CAD of breast MRI is the definition of the breast area, which is enclosed by air and the chest wall. Several methods have shown good results to detect the anterior tissue-air boundary[8, 9]. However, detection of the posterior boundary between the breast and the chest wall, which is in breast MRI images comprised of the pectoral muscle and sternum, has not been fully resolved[9]. Current methods have reported large challenges to delineate the muscle border correctly in patients with dense breasts (ACR 4) because the contrast between muscle and glandular tissue is poor[10–12].

The chest wall is typically delineated using semi-automated computer assisted methods[13, 14] or automated methods that result in a roughly estimated chest volume, used by for example CADstream (Merge Healthcare Inc., Chicago, IL) and DynaCAD (Invivo, Gainesville, FL). Several fully automated detailed methods have also been reported. These detailed methods can be divided in two groups, knowledge-based methods and deep learning-based methods. Knowledge-based methods use intensity operations and gradient signs[15, 16], edge properties[8, 17–20] or a-priori atlases[9, 10]. Deep learning-based methods for chest wall segmentation have used artificial neural networks in the form of convolutional neural networks[9, 15, 21]. The performance of these methods is difficult to compare as for each method results have been reported for different data sets,

which vary widely in the number of ACR 4 images included. The largest ACR 4 data set on which a knowledge-based method has been evaluated contained 55 cases[17], while the largest data set on which a deep learning-based method has been evaluated contained 15 ACR 4 cases[21]. The reported Dice similarity coefficient (DSC) of deep learning-based methods to segment the chest wall in extremely dense breasts is 0.921[21]. The performance of knowledge-based methods ranges from 0.944 to 0.96[9, 17, 19, 20]. A direct comparison of the two approaches using a large MRI dataset of ACR 4 breast would shed light on the advantages and pitfalls of both approaches.

The aim of this study is to compare a knowledge-based and a deep learning-based approach for segmentation of the chest wall in MR image of extremely dense breasts. Both methods were optimized and validated using an identical large dataset. Using a large series of MR images, we pursued to minimize selection bias and cover the variety of ACR-4 breast types. The secondary objective of this study was to test the effectiveness of various quantitative metrics to assess chest wall segmentation results in terms of clinical relevance.

2.3 Materials and Methods

2.3.1 Study population

MRIs were collected from 115 randomly selected participants in the Dense Tissue and Early Breast Neoplasm Screening (DENSE) trial, who were examined in the University Medical Center Utrecht, Utrecht, the Netherlands. The DENSE trial has been described in detail elsewhere[2]. In short, this multicenter randomized controlled trial investigates the additional value of MRI screening in Dutch women with extremely dense breasts (i.e. ACR4). Written informed consent was obtained from all patients before MRI screening. The trial was approved by the Dutch Minister of Health, Welfare and Sport (2011/19 WBO, The Hague, the Netherlands). The age of the participants ranged from 49.5 to 75.2 years with an average of 56.6 years. None of the selected participants had a lesion suspected of being malignant.

2.3.2 MR Imaging

This study used a random subset of 115 T1-weighted MRI breast scans, acquired in the UMC Utrecht. Each scan consisted of 179 slices. All participants were

scanned in prone position using a Philips Achieva 3 Tesla MR scanner (Philips Healthcare, Best, The Netherlands). The image data consisted of high-spatial resolution transversal images, obtained with a 3D sequence using a dedicated phased-array bilateral breast coil (Philips SENSE-Breast7TX receive coil) with a repetition time of 4.95 ms, an echo time of 1.87 ms and a 10° flip angle. Single slice dimensions were 560 x 560 pixels, the field of view was 360 x 360 mm² and the in-plane resolution was 0.64 x 0.64 mm² with a slice thickness of 1.00 mm. To train and validate both automated segmentation methods, the dataset was randomly split in a training set of 36 image sets and a validation set of 79 image sets. Manual reference segmentations of the chest wall were obtained in all datasets and used as ground truth for optimization and validation. The segmentation was performed by contouring in 2D transverse images by a Technical Physician (EV) under supervision of a breast radiologist (SW) who had 6 years' experience with breast MRI. Interactive tools such as Livewire[22] and freehand contouring were used, available in MeVisLab (version 3.0, MeVis Medical Solution AG, Bremen, Germany).

2.3.3 Methods

Two methods for segmentation of the chest wall were developed for the purpose of this study: A knowledge based method (Section Knowledge-based chest wall segmentation) and a deep learning-based method (Section Deep learning based chest wall segmentation). For both methods the same image-preprocessing step was used (Section Pre-processing), the performance of the methods was compared (Section Evaluation) and the effect of combination of the methods was reviewed.

Pre-processing

Prior to segmentation, data was preprocessed. Preprocessing started with the automated definition of a rectangular region of interest (ROI), containing the area between 1 cm anterior of the breast tissue and 5 cm posterior of the intermammary cleft. First, the MRI volume was separated into foreground and background voxels using Otsu's method[23]. Then, three landmarks were automatically detected in the resulting binary images (Figure 2.1). Landmark 1 corresponded to the most anterior tissue in the image dataset, often the nipple of one of the breasts. Landmark 2 was the corresponding landmark at the same transversal

image slice. Landmark 3, was the most posterior air-tissue boundary between the two detected landmarks, denoting the intermamillary cleft. For all subsequent analysis, the MRI volume was cropped to this ROI.

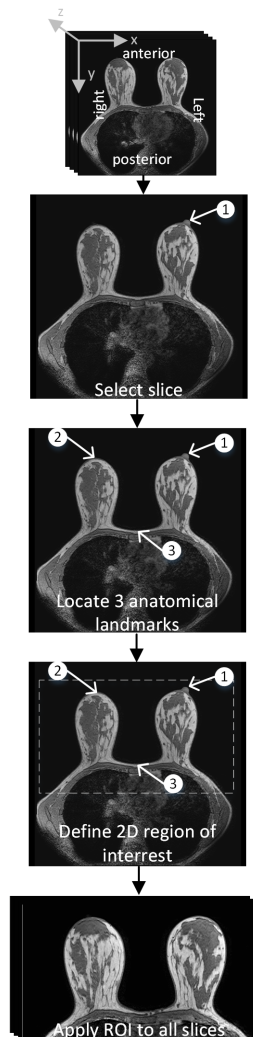


Figure 2.1: The definition of the automatically established ROI. Landmark 1 is the location most anterior tissue, landmark 2 the location of the most anterior tissue in the contra lateral part of the same image slice and landmark 3 the position of the intermamillary cleft. The ROI was defined around the landmarks and applied to all slices of the dataset.

Knowledge-based chest wall segmentation

A knowledge-based method was developed to find the curve dividing the chest wall and breast in each image. The method combined heuristics and dynamic programming[17, 20, 24] in two steps. First, a cost image was formed. Second, a path-finding algorithm was used in the cost image to trace the border of the chest wall. The total cost of a path is the sum of cost values along the path in the underlying image. A high cost is assigned to locations where the chest wall is unlikely to be present (e.g., inside the lungs and anterior of the breast). A low cost is assigned to locations where a clear edge contrast is present. Using this approach, the shortest path between the two bottom corners of each transversal slice is forced to go around the lungs and favor a more posterior path which follows a clear edge as much as possible. All steps of this automated segmentation were performed using MATLAB (v R2015a; Mathworks, Natick, MA) running on a desktop PC (Intel Xeon CPU 3.50GHz, 16GB RAM).

Each image was transformed into a cost image, c (Figure 2.2). In c , each voxel value is inversely proportional to the likelihood that the border between breast and chest is present at that location. Four three-dimensional image layers ($L_1 - L_4$) formed the cost image. Each layer used image properties to score the cost at each voxel. The cost function was defined as:

$$c(x, y, z) = L_1(x, y, z) * (\alpha * L_2(x, y, z) + L_3(x, y, z) + \gamma * L_4(x, y, z))^2 + (\delta * f'(x, y, z)) + 0.1 \quad (2.1)$$

Where L_1-L_4 represent different image properties, x , y and z are the coordinates of each voxel and α , γ and δ are weighting factors to favor or penalize a specific layer. Image f' is the cropped image where voxel intensities were normalized between 0 (no intensity) and 100 (maximal intensity). Physical path length was penalized by addition of 0.1 to all voxels in the cost image.

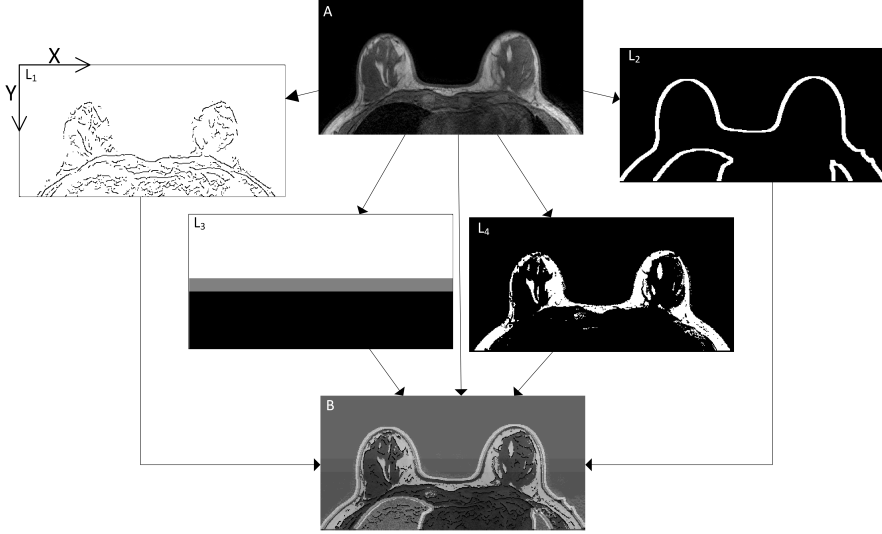


Figure 2.2: Overview of layers used to form the Cost image, showing input image, A, the layers of the cost image, L_1 to L_4 and the resulting cost image, B. L_1 contains the edges present in the A, L_2 adds information on borders, L_3 about vertical position relative to the intermamillary cleft, L_4 about glandular fatty tissue distribution

Layer L_1 contains the edges resulting from contrast differences present in the image f' . A part of the border of the chest wall in T1-weighted MR images is located between low image-intensity muscle tissue and high image-intensity fatty tissue or low image-intensity glandular tissue.

$$L_1(x, y, z) = \begin{cases} 1 - E(x, y, z) & \text{where } (\frac{\delta f'}{\delta y}) < 0 \\ 1 & \text{otherwise} \end{cases} \quad (2.2)$$

Thus, layer L_1 is a binary image where voxels at detected edges at positions with negative y gradient (anterior-posterior direction) are assigned value 0 and all other voxels value 1. E is the binary result of the Canny edge detection filter[25] using hysteresis thresholds $T_l = 0.0125$ and $T_U = 0.0625$ and $\delta = \sqrt{2}$ for the Gaussian smoothing applied to the normalized image f' . The hysteresis thresholds were chosen relatively low to include more potential edges.

Layer L_2 also is a binary three dimensional volume which penalizes the edges, and voxels anterior to the edges found at the transition between tissue and air:

$$L_2(x, y, z) = G(x, y, z) \oplus SE_2 \quad (2.3)$$

where

$$G(x, y, z) = A(x, y, z) \oplus SE_1 - A(x, y, z) \quad (2.4)$$

Where A is the complement of the tissue segmentation obtained using threshold T_0 (which is also used during preprocessing). G is the inner border of the volume segmented as tissue in A , obtained by dilation (\oplus) of A with a $3 \times 3 \times 1 \text{ mm}^3$ structuring element, SE_1 . L_2 is the dilation result of G with structuring element SE_2 . Element SE_2 spans $5.5 \times 3.0 \times 1.0 \text{ mm}^3$ with the origin in the center of the posterior side of the element. Layer L_2 was weighed by an arbitrarily large factor (α), yielding large cost for the edges found at the transition between tissue and air.

The third layer, L_3 , penalized voxels located anterior of the intermammillary cleft, (Figure 2.3) Voxels in rows located between 15 mm and 30 mm anterior from the intermammillary cleft were penalized with weighting factor β , while voxels in rows located more than 30 mm from the intermammillary cleft were penalized with weighting factor α .

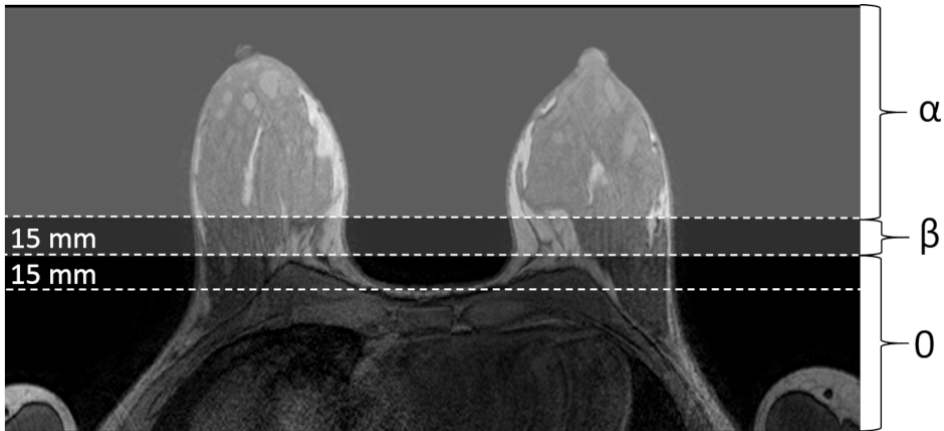


Figure 2.3: Layer L_3 illustrated as an overlay on an MR image of the breast, weighting factors are indicated on the right of the image.

The fourth layer, L_4 , exploits the fact that the fatty tissue typically has higher image intensity on T1-weighted images than other anatomical structures. The image was segmented into three pixel-value classes: low, intermediate, and high

image intensity, using fuzzy-c means clustering[26] with a fuzziness factor of 2. In L_4 , all voxels that are part of the high image-intensity class were set to 1, all other voxels were set to 0.

The cost image, c , was used for tracking the path that accumulates the smallest sum of cost values from the left to the right posterior corner. Dijkstra's algorithm[27] was used for this purpose. The path finding algorithm was applied to all transversal slices of the cost image separately resulting in one path for each slice. All paths combined resulted in an irregular three-dimensional surface, P . The surface of this volume was labeled with value 1 and all other voxels in the image with 0.

To remove irregularities from the surface, a 2D surface topography map, H , was formed from surface P . In H , each pixel was assigned a value equal to the shortest distance of the surface voxels to the posterior side of the image matrix. A median filter using a kernel of $15 \times 15 \text{ mm}^2$ was subsequently applied to the topography map to remove large fluctuations in height. The median filtered surface topography map, was converted back to a surface M in 3D, where the surface was labeled with value 1 and the other voxels with value 0. Finally, a new cost image was composed from surface M , the binary result of the Canny edge detection filter, E , and the first cost image, c :

$$c_2(x, y, z) = \begin{cases} 0.1 & \text{if } M(x, y, z) = 1 \text{ or } E(x, y, z) = 1 \\ 0.01 & \text{if } M(x, y, z) = 1 \text{ and } E(x, y, z) = 1 \\ \epsilon * c(x, y, z) & \text{otherwise} \end{cases} \quad (2.5)$$

Here, ϵ is a weighting factor, set to value 100. Cost image, c_2 , was used to find the final chest wall segmentation. First, the path in the middle transversal slice was tracked. This was repeated slice by slice, taking into account the path found in the adjacent slice as follows: Before the path in a next slice is tracked, the corresponding transversal slice in cost image c_2 was updated to prevent irregular surface outcomes. All voxels in c_2 located 4 mm or more from the path in the adjacent slice were maximally penalized using factor α , which arranges that the resulting path will not deviate more than 4 mm from the path found in the previous segmented adjacent slide.

Optimal results, in the training data, were achieved using empirically determined weighting factors $\alpha=1e16$, $\beta=10$, $\gamma=10$, $\delta=2$ and $\epsilon=100$. Robustness of the weighting factors was tested in two ways. First, all possible permutations (repetition allowed) of weighing factors 2, 10 and 100 for $\alpha, \beta, \gamma, \delta$ and ϵ were evaluated to

segment the training data. This confirmed the chosen selection of weighting factors which resulted in a median Dice similarity coefficient (DSC)[28] of 0.985. The DSC is a measure for the overlap of the segmented breast volume and the ground truth. Median DSC ranged from 0.910 to 0.985 for all other possible permutations of the weighting factors (Supplemental material: Robustness of the KBM). Secondly, uniform random noise, with a maximum up to 100% of the weighing factor, was added to the weighting factors. This did not lead to significantly different results ($p = 0.19$) in the training data with a median DSC of 0.983 compared to a DSC of 0.985 in noiseless images.

Deep learning based chest wall segmentation

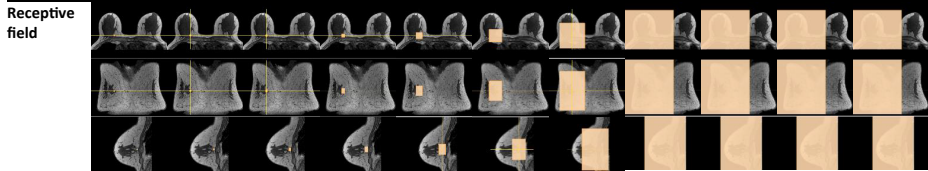
The second proposed method uses a dilated convolutional neural network (DCNN) to segment the chest wall. For this method, we defined chest wall segmentation as a two-class segmentation problem, where the DCNN should predict a label 0 for voxels anterior to the chest wall and a label 1 for voxels posterior to the chest wall. In contrast to a non-dilated CNN, a DCNN stacks layers with increasing rates of dilation, i.e. increasing spacing between kernel elements.

The receptive field is the part of the image taken into account to predict class probabilities for a voxel. By linearly increasing the dilation rate from layers 2 through 7, the receptive field grows exponentially to a final width of 131x131 voxels. However, the number of parameters in each layer stays the same. The proposed DCNN provide a receptive field of 259x259 in only 9 layers (Table 1). Without dilation, a receptive field of this size requires at least 129 layers using 3x3 convolutional kernels. The dilation rate does not affect the number of parameters in a kernel, but it does lower the number of layers. Hence, a large amount of context can be taken into account with a low number of trainable parameters, and hence a reduced chance of overfitting[29, 30].

For each layer of the DCNN used in this study, Table 2.1 lists the number of convolution kernels, the convolution kernel size, the convolution kernel dilation, and the number of parameters. In addition, the receptive field at each layer is listed, i.e. the part of the image that is taken into account to predict a value for a single voxel. The dilation rate is linearly increased from 1 to 64 between layers 2 and 8. This means that at layer 2, there is no spacing between kernel elements, and at layer 8, kernel elements are spaced 63 pixels apart. By linearly increasing the dilation rate the receptive field grows exponentially to a final width of 259x259 pixels. However, the number of parameters in each layer stays the same. To

preserve translational equivariance, the kernel stride is 1 in all cases. Layers 1 to 10 are each followed by a rectified linear unit (ReLU), while layer 11 is followed by a sigmoid function. No skip connections were used in the network.

Table 2.1: The convolutional neural network architecture used in this study. For each layer, the convolution kernel size, the rate of dilation, the receptive field, the number of output channels and the number of trainable parameters are listed. The kernel stride is 1 in all cases. Figures in the top row illustrate the receptive field at each layer shown in orange.



Layer	1	2	3	4	5	6	7	8	9	10	11
Convolution	3 x 3	3 x 3	3 x 3	3 x 3	3 x 3	3 x 3	3 x 3	3 x 3	3 x 3	1 x 1	1 x 1
Dilation	1	1	2	4	8	16	32	64	1	1	1
Field	3 x 3	5 x 5	9 x 9	17 x 17	33 x 33	65 x 65	129 x 129	257x257	259 x 259	259 x 259	259 x 259
Channels	32	32	32	32	32	32	32	32	32	192	3
Parameters	320	9248	9248	9248	9248	9248	9248	9248	9344	6912	579

Given a 2D input sample of 259x259 pixels, the DCNN in Table 1 will predict a single value for the center pixel. As the DCNN only uses valid convolutions, no values will be predicted for the 129 border pixels in each direction. However, any image larger than 259x259 pixels can also be processed, resulting in a prediction for the voxels in the center of that image. We use this principle during training and testing. During training, we provided the DCNN with 409x409 pixel samples and reference predictions for the 151x151 pixels in the center (Figure 2.4). During testing, we provided full-size 2D images to the DCNN. To accommodate for the loss of border pixels, 3D volumes were padded with voxels in each direction prior to training or testing.

A single DCNN was trained to segment 2D images in three orthogonal directions: transversal, sagittal, or coronal. For this, the 36 datasets in the training group were stratified into two groups: a training set containing 35 data sets and a validation set containing 1 data set. The latter was used for hyperparameter optimization. The network was trained for 100,000 iterations using the Adam optimizer [31] with a learning rate of 0.0001. During each training iteration, a mini-batch consisting of 10 images of size 409x409 pixels and corresponding reference segmentations of 151x151 pixels (Figure 2.4) was randomly selected from the training set. The DCNN was trained to minimize the Dice loss [32].

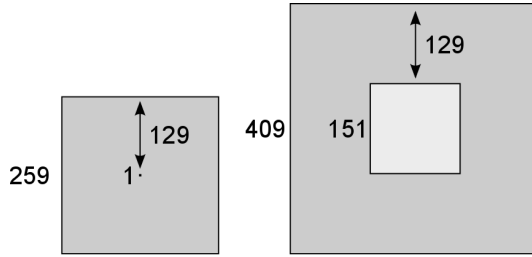


Figure 2.4: Given an input image of 259x259 pixels, the DCNN will predict a single value. During training, the DCNN is provided with 409x409 pixel input samples, and a prediction is made for the 151x151 center pixels, with no loss of resolution.

$$Diceloss = 1 - \frac{2 \sum_i (As_p_i * Ms_i)}{\sum_i (As_p_i * Ms_i)} \quad (2.6)$$

where As_p is the resulting 3D probability volume, Ms is the binary 3D volume of the manual segmented chest and i is iterating over all voxels. The DCNN was implemented in Python with PyTorch and experiments were performed using an NVIDIA Titan X GPU with 12GB RAM.

During testing, the trained DCNN was directly applied to all 2D images along the three principal axes of the test image to obtain three 3D probability volumes Figure 2.5. These were averaged and thresholded at $p=0.5$ to obtain a binary prediction. Finally, to obtain the surface delineating the chest wall, the largest component with label 1 (i.e., posterior to the chest wall) was identified in the binary prediction. A morphological erosion was applied to this component and the resulting mask was subtracted from the component so only the boundary voxels remain without affecting the border itself.

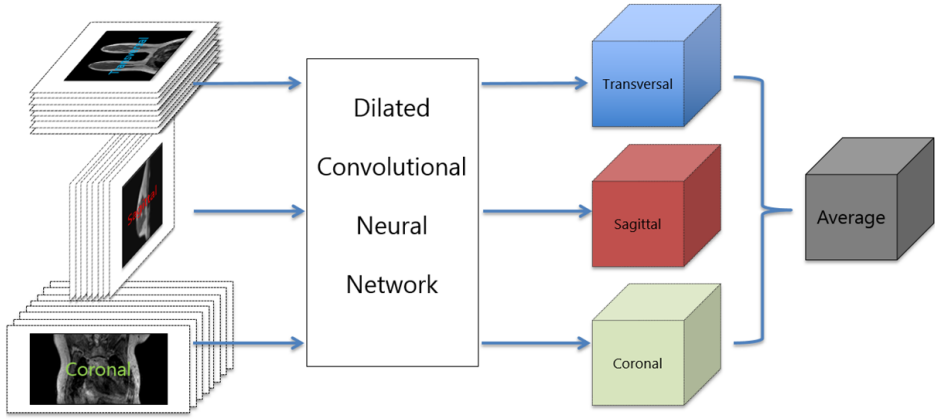


Figure 2.5: The dilated neural network segments each dataset along the three principal axes resulting in three equal volumes with a segmentation result of the chest volume. The three resulting class probabilities were averaged and thresholded to obtain the final segmentation.

Evaluation

The automated segmentation method was validated using qualitative and quantitative metrics. Seventy-nine data sets were automatically segmented for validation. Quantitatively, the manual and automatic segmentation results were compared using the total 3D chest volume and slice-by-slice 2D chest volume using four validation metrics: the Dice similarity coefficient (DSC) [28][28], false positive fraction (FPF) or over-segmentation[17], false negative fraction (FNF) or under-segmentation[17] and Hausdorff distance (HD). The quantitative results of both segmentation methods were compared using the Wilcoxon signed rank test, a p-value of less than 0.05 was considered statistically significant. The DSC metric was calculated using equation 2.7 characterizing an overlap measure, i.e., the area agreement, between the automatically (A_s) and manually (M_s) obtained chest volumes delineated by the found chest wall border and posterior edge of the ROI.

$$DSC = \frac{2(A_s \cap M_s)}{(A_s + M_s)} \quad (2.7)$$

The FPF and FNF were calculated using equation 2.8 and equation 2.9

$$FPF = \frac{A_s \setminus M_s}{A_s \setminus M_s + A_s \cap M_s} \quad (2.8)$$

$$FNF = \frac{Ms \setminus As}{MS \setminus As + As \cap Ms} \quad (2.9)$$

Where \cap denotes the intersection of the set of volume pixels and \setminus the difference operator. The fourth metric to score the automatically obtained segmentations was the HD metric. The HD reflects the maximal Euclidian distance between manually and automatically delineated borders. The HD is generally sensitive to outliers, therefore we used the quantile method proposed by Huttenlocher et al. [33]. According to the Hausdorff distance quantile method, the HD is defined to be the q^{th} quantile of distances, instead of the maximum. In this study, we selected the 95th percentile, HD95th

In addition to the quantitatively scoring all segmented slices were scored qualitatively by a radiologist. For this purpose, segmentation errors were divided in four categories:

1. Correctly delineated
2. Under-segmentation: chest wall is segmented as breast tissue.
3. Mild over-segmentation: soft tissue other than breast tissue is segmented as chest wall, or deviation to target is smaller than 2 mm.
4. Severe over-segmentation: breast tissue is segmented as chest wall.

Any form of over-segmentation, where breast tissue was segmented as chest, was considered to be worse than under segmentation because it masks breast tissue and may thus cover breast lesions. Based on the BIRADS atlas[34], where breast foci have a maximal diameter less than 5 mm and breast lesions have a maximal diameter of at least 5 mm, we set the threshold between mild and severe over segmentation at 2 mm to minimize the chance of missing a lesion mass when mild over-segmentation was present. The qualitative results of both methods were compared using the McNemar chi squared test where p-value smaller than 0.05 is considered significant. Furthermore, associations between quantitative results and qualitative results were shown using the Wilcoxon signed-rank test.

2.4 Results

Quantitatively, no significant difference was present between both methods according to the DSC and HD metric. Nonetheless, the FPF of the KBM was sig-

nificantly lower compared than that of the DLM. Both FNF and HD95 were significantly lower in the DLM compared to the KBM (Figure 2.6). In other words, the KBM outperformed the DLM in terms of FPF, but the DLM outperformed the KBM in terms of FNF and HD95.

Qualitatively compared, the DLM performed significantly better than the KBM according to the McNemar chi square test, $p < 0.01$. The success rate of the DLM was higher compared to the success rate of the KBM, 0.82 versus 0.78 respectively (Table 2.2). In 7.9% of the slices both methods were not successful, in other words 92.1% of the slices were segmented correctly by one of the methods. We found that the qualitative analysis reflected the numbers found in the quantitative analysis. Slices that were scores as category 1 had the highest DSC, while slices scored as category 3 and 4 had the highest FPF (Figure 2.7). Upon closer inspection, we found that the DLM mostly had problems finding the correct chest wall in MR images of rare anatomy, for example where glandular tissue did continue deep into the axilla. Conversely, the tendency of the KBM to find a minimum cost path sometimes resulted in unwanted shortcuts. These occurred when the pectoral muscle had irregularities, e.g. in or near the shoulder where the detected border was located posterior of the chest wall border (Figure 2.8).

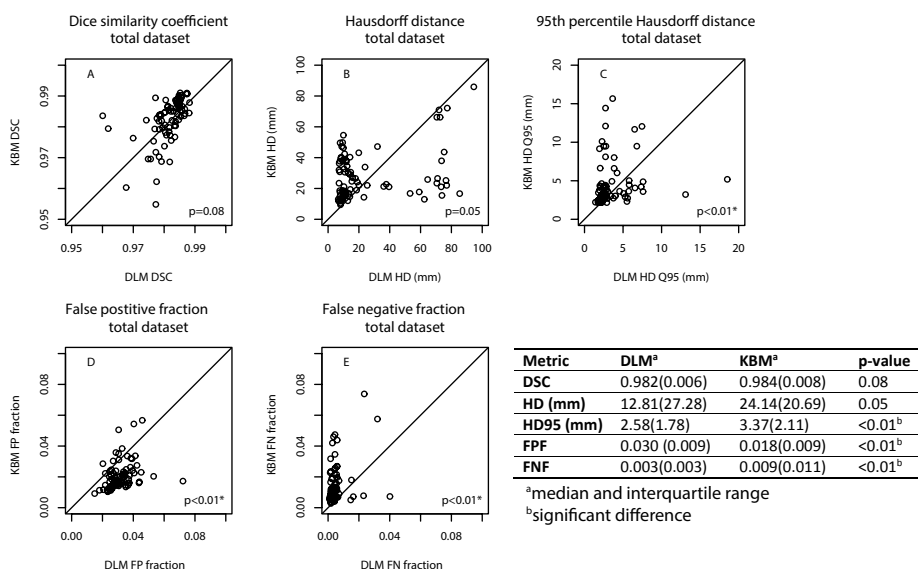


Figure 2.6: Quantitative results of automated segmentation methods compared to the ground truth. Result of statistical comparison between methods using the Wilcoxon signed rank test are shown in the bottom right corner of each graph. The table shows a summary of the quantitative results.

Table 2.2: Shows an overview of the rating of the radiologist of each segmented slice for both methods. Category 1 is successful segmented, category 2 is for under segmentation, category 3 mild over segmentation and category 4 severe over segmentation. As shown in the table 9596 (67.9%) slices were segmented correctly by both methods and 13029 (92.1%) slices were segmented correctly by at least one of the proposed segmentation methods.

	DLM category 1	DLM category 2	DLM category 3	DLM category 4	Total
KBM category 1	9596	648	413	430	11087
KBM category 2	929	303	120	124	1476
KBM category 3	515	44	120	99	778
KBM category 4	498	66	104	132	800
Total	11538	1061	757	785	14141

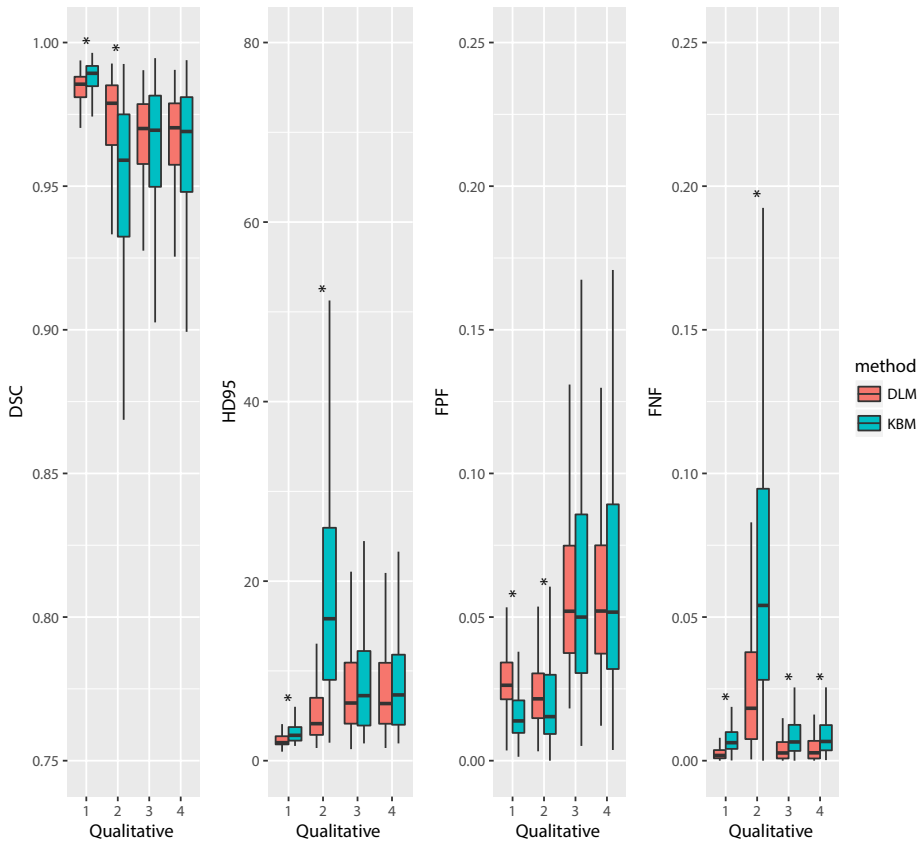


Figure 2.7: Relation between slice by slice quantitative and qualitative scoring. On the x-axis, the 4 categories of quantitative scoring and the qualitative scoring on the y-axis. A significant difference in performance between DLM (red) and KBM (blue) is shown by the *.

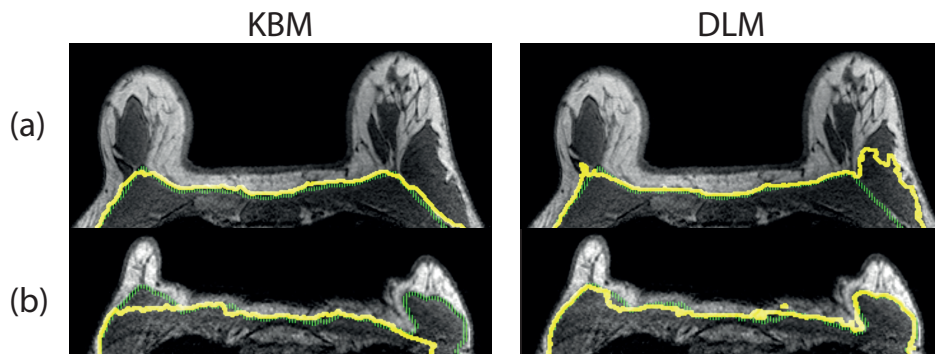


Figure 2.8: Worst-case segmentations by both approaches. Left the result of the KBM (yellow), right the result of the DLM (yellow), A) transversal slice of dataset where the lowest DSC (0.955) occurred between manual segmentation (green and dashed) and result of DLM (yellow). B) transversal slice of dataset where the lowest DSC (0.960) occurred between manual segmentation (green and dashed) and result of KBM (yellow). Segmentation results were dilated for visibility.

2.5 Discussion

This study shows that an optimized knowledge-based method and deep learning method can trace the chest wall border in 79 independent MRI datasets (i.e., not previously seen) of extremely dense breasts (i.e., ACR class 4). Both methods were qualitatively and quantitatively evaluated in data set consisting of 79 new screening MR images of extremely dense breasts. For the DSC, no significant difference was found between the two methods. Although the KBM showed better performance in terms of FPF, the DLM outperformed the KBM in terms of FNF and HD95.

A number of studies on chest wall segmentation have been published that describe and evaluate the results in quantitative terms. Only few studies[10, 12, 17] described the type of segmentation error, while none described the location of the error. These are important aspects, because some errors may lead to clinically unacceptable results while others will have negligible clinical impact. The quantitative results of other methods are summarized in Table 2.3. As shown, both methods presented in this study perform at least as well or better than previously published methods, although the methods are not directly comparable due to the use of different datasets. For a fair comparison, we additionally implemented a state-of-the-art knowledge based method and a deep learning based method to segment the same test data. We chose the KBM method of Milenkovic[20] because it is the best performing method reported in the literature for extremely

dense breast images without fat suppression. In addition we chose the widely used DLM method U-Net[35]. Performances of both methods are shown in Table 2.3.

Table 2.3: Comparison of the achieved metric values (\dagger mean and standard deviation or \ddagger median and interquartile range) with those found in the literature. When authors split their results by ACR category only the results of the segmented ACR 4 cases are mentioned. Methods from literature used to segment the same test data are indicated by an *.

Author	Number of MRI datasets	DSC	FNF	FPF	HD95 (mm)	HD (mm)	ACR 4	Method
Fooladivanda[17] \dagger	55	0.946 (0.03)	0.035 (0.021)	0.072 (0.042)	NA	NA	Yes	KB
Gallego Ortiz[10] \ddagger	409	0.88 (0.05)	0.11 (0.05)	0.13 (0.07)	NA	NA	No	KB
Gubern Merida[12] \ddagger	50	0.94 (0.03)	0.04 (0.02)	0.07 (0.06)	NA	NA	No	KB
Milenkovic[20] \dagger	11	0.949 (0.018)	NA	NA	NA	NA	Yes	KB
Wu[8] \ddagger	14	0.944 (0.024)	NA	NA	NA	NA	Yes	KB
Jiang[19] \dagger	8	0.96 (0.011)	NA	NA	NA	NA	Yes	KB
Wei[18] \ddagger	99	0.960(0.017)	0.02(0.02)	0.01(0.01)	NA	NA	No	KB
Dalmis[21] \ddagger	15	0.921 (0.03)	NA	NA	NA	NA	Yes	ML
Milenkovic[20]* \ddagger	79	0.956 (0.026)	0.012 (0.006)	0.072 (0.056)	8.42 (4.43)	34.08 (18.63)	Yes	KB
Ronneberger[35]* \ddagger (U-Net)	79	0.983 (0.004)	0.003 (0.002)	0.029 (0.007)	2.21 (0.75)	11.93 (43.08)	Yes	ML
Proposed methods								
DLM \ddagger	79	0.982 (0.006)	0.003 (0.003)	0.030 (0.009)	2.58 (1.78)	12.81 (27.28)	Yes	ML
KBM \ddagger	79	0.984 (0.008)	0.009 (0.011)	0.018 (0.009)	3.37 (2.11)	24.14 (20.69)	Yes	KB

In problems with small amounts of training data, the large number of trainable parameters of a U-Net (31 million) could increase the risk of overfitting compared to a DCNN which is using far fewer trainable parameters (82 thousand). However, in this study the proposed DLM is on-par with U-Net. The knowledge based method of Milenkovic performs well, but has significantly lower DSC, and significantly higher FNF, FPF and HD95 compared to both proposed methods.

It should be noted that the presented DSC values are of the chest volume were all DSC values reported by other authors are of the breast volume. This choice was made because the breast volume is also enclosed by the border between skin and air whereas the chest volume is only enclosed by the border of the ROI and found chest wall. Hence, all results were solely based on the chest wall segmentation and not on segmentation errors of the skin-air border.

To the best of our knowledge, no studies have yet systematically compared optimized knowledge-based heuristic methods and artificial-intelligence methods for this problem. This study shows that both presented approaches have advantages

and disadvantages. From a practical point of view, deep learning runs significantly faster and yields more robust performance in terms of FNF and HD95, while FPF metric showed a better performance for the KBM. We expect that false positive results have more severe impact when used in computer aided diagnosis because the chance of missing malign lesions increases, while false negative results will never remove breast lesions. However for fibroglandular volume or breast parenchyma enhancement measurement (BPE) false negative segmentation results, which are less present at the DLM, can result in volume over estimation or wrong BPE values. The complementary nature of both methods resulted in only 7.9% of the slices that were not fully correct segmented when considered jointly.

Existing methods may show difficulties tracing the chest wall border due to the lack of contrast between glandular tissue and chest wall tissue[10–12] or they perform worse in extremely dense cases compared to segmentation in images of less dense breast[8]. This study showed that the presented methods can trace the chest wall border in 79 extremely dense breast MRI in an independent test set. Since extremely dense breasts are considered to be the most difficult cases for automatic segmentation of the chest wall. It is reasonable to assume the performance of the KBM will be consistent or better in MR imaging of breasts with less glandular tissue. However, for the DLM, training data with less glandular tissue should also be present in the training data before the method is expected to perform comparable.

Chest wall segmentation often is a preprocessing step in automated analysis of breast imaging, for example to measure breast volume and glandular tissue volume or prior to computer aided detection of lesions inside the breast. With the KBM, where the shortest path could result in unwanted shortcuts, the number of false positives is reduced but it increased the amount of false negatives. The DLM suffers from larger false positives fractions but performs better as a whole. Therefore we prefer to use the KBM when the aim is to detect breast lesions, because there is less chance a lesion is hidden due to a false positive chest wall segmentation, but when the aim is to measure breast volumes we prefer the DLM. As expected there was a relation between the DSC metric and the scoring of the radiologist. Also the relation between qualitative false positive categories (3 and 4) and false negative category (2) and quantitative metrics FPF and FNF was as expected.

As described by Milenkovic et al.[20], a risk of the dynamic programming ap-

proach (KBM) is the success of the first slice being segmented: when this segmentation is incorrect, the error will propagate through to the adjacent slice. In this study we selected the middle transverse slice, because this slice is near iso-center of the MRI scanner where the signal to noise ratio is optimal. Alternative locations to start this method are slices more superior or inferior, where the amount of glandular tissue is reduced. The parameter settings in the KBM were determined empirically by visual inspection of the layers. In this work, we used a dilated CNN (DCNN) for segmentation. This architecture has previously shown excellent performance on medical image analysis tasks[29, 30, 36]. The DCNN was trained to segment 2D slices in the transversal, sagittal, and coronal images. We also evaluated a DCNN that was trained to only segment transversal, coronal or sagittal slices. This led to significantly lower performance (Supplemental material 2: Variations on DLM), indicating that there is useful information in a combination of the planes. To further investigate this, we extended the 2D DCNN to 3D by using 3D convolutional kernels. However, this required compromising on the size of the receptive field to accommodate limitations in available RAM on a typical GPU. Therefore, predictions for a voxel depended on a receptive field of $131 \times 131 \times 9$ voxels centered at that voxel. This network performed on par with the proposed DLM for all metrics except the HD95 metric, where it had a significant lower performance ($p < 0.01$) (supplemental material 2: Variations on DLM). In future work, we may explore other neural network architectures, such as those with multi-scale patch-based networks[37] or ensembles of different architectures[38] for improved results. A study limitation is that all data were acquired in the same hospital. It is known that the variation in MR images quality can be substantial, therefore, for future research we advise to increase the variation by using images acquired using MRI systems from different vendors and from different hospitals. It is conceivable that the accuracy of segmentation results would be increased with more data, or the methods become more versatile for different protocols. To achieve this goal, the KBM may need to adapt its cost functions to different protocols, while the parameters in the DLM may need to be fine-tuned to generalize across protocols. In most cases, for the DLM this may mean retraining of the DCNN with the same hyperparameters. However, if increased complexity of training data can no longer be accurately represented by the same number of parameters in the network, the DCNN architecture may need to be adjusted and new hyperparameters may have to be used. All methods in this study were developed using a training and validation set, and evaluated on a separate hold-out test set. This paradigm was chosen

over cross-validation, because knowledge-based methods cannot be developed in a cross-validated fashion and our aim was to compare methods on exactly the same test set. Finally the effects of inter- and intra-observer variability in obtaining the manual ground truth segmentation are unknown in this study and could give more insight about the differences in performances of the methods.

2.6 Conclusion

We developed two automated methods for segmentation of the chest wall in MR images of extremely dense breasts. Both methods were evaluated on an independent dataset of 79 MR examinations, and showed a good performance. Both methods have their strengths and weaknesses. Hence, we consider that the KBM is more suitable for methods where the aim is breast lesion detection and the faster DLM is preferable when measuring breast volumes, which is important when determining breast density.

References

- [1] L. Kauhava et al., "Lower costs of hospital treatment of breast cancer through a population-based mammography screening programme," *Eur J Public Health*, vol. 14, no. 2, pp. 128–33, 2004.
- [2] M. J. Emaus et al., "Mr imaging as an additional screening modality for the detection of breast cancer in women aged 50-75 years with extremely dense breasts: The dense trial study design," *Radiology*, vol. 277, no. 2, pp. 527–37, 2015.
- [3] N. F. Boyd et al., "Mammographic density and the risk and detection of breast cancer," *N Engl J Med*, vol. 356, no. 3, pp. 227–36, 2007.
- [4] K. Kerlikowske, "The mammogram that cried wolfe," *New England Journal of Medicine*, vol. 356, no. 3, pp. 297–300, 2007.
- [5] E. R. Price et al., "The california breast density information group: a collaborative response to the issues of breast density, breast cancer risk, and breast density notification legislation," *Radiology*, vol. 269, no. 3, pp. 887–92, 2013.
- [6] V. A. McCormack and I. dos Santos Silva, "Breast density and parenchymal patterns as markers of breast cancer risk: a meta-analysis," *Cancer Epidemiol Biomarkers Prev*, vol. 15, no. 6, pp. 1159–69, 2006.
- [7] E. Warner et al., "Systematic review: Using magnetic resonance imaging to screen women at high risk for breast cancer," *Annals of Internal Medicine*, vol. 148, no. 9, pp. 671–679, 2008.
- [8] S. Wu et al., "Automated chest wall line detection for whole-breast segmentation in sagittal breast mr images," *Med Phys*, vol. 40, no. 4, p. 042301, 2013.
- [9] A. Gubern-Merida, M. Kallenberg, R. Marti, and N. Karssemeijer, "Segmentation of the pectoral muscle in breast mri using atlas-based approaches," *Med Image Comput Comput Assist Interv*, vol. 15, no. Pt 2, pp. 371–8, 2012.
- [10] C. G. Ortiz and A. L. Martel, "Automatic atlas-based segmentation of the breast in mri for 3d breast volume computation," *Med Phys*, vol. 39, no. 10, pp. 5835–48, 2012.
- [11] M. Lin et al., "Template-based automatic breast segmentation on mri by excluding the chest region," *Med Phys*, vol. 40, no. 12, p. 122301, 2013.
- [12] A. Gubern-Merida et al., "Breast segmentation and density estimation in breast mri: a fully automatic framework," *IEEE J Biomed Health Inform*, vol. 19, no. 1, pp. 349–57, 2015.
- [13] C. Klifa et al., "Magnetic resonance imaging for secondary assessment of breast density in a high-risk cohort," *Magn Reson Imaging*, vol. 28, no. 1, pp. 8–15, 2010.
- [14] D. Kontos et al., "A comparative study of volumetric breast density estimation in digital mammography and magnetic resonance imaging: Results from a high-risk population," *Medical Imaging 2010: Computer - Aided Diagnosis*, vol. 7624, 2010.

- [15] G. Ertas et al., "Breast mr segmentation and lesion detection with cellular neural networks and 3d template matching," *Computers in Biology and Medicine*, vol. 38, no. 1, pp. 116–126, 2008.
- [16] T. Twellmann, O. Lichte, and T. W. Nattkemper, "An adaptive tissue characterization network for model-free visualization of dynamic contrast-enhanced magnetic resonance image data," *IEEE Transactions on Medical Imaging*, vol. 24, no. 10, pp. 1256–1266, 2005.
- [17] A. Fooladivanda, S. B. Shokouhi, and N. Ahmadinejad, "Localized-atlas-based segmentation of breast mri in a decision-making framework," *Australas Phys Eng Sci Med*, vol. 40, no. 1, pp. 69–84, 2017.
- [18] D. Wei et al., "Three-dimensional whole breast segmentation in sagittal and axial breast mri with dense depth field modeling and localized self-adaptation for chest-wall line detection," *IEEE Transactions on Biomedical Engineering*, pp. 1–1, 2018.
- [19] L. Jiang et al., "Fully automated segmentation of whole breast using dynamic programming in dynamic contrast enhanced mr images," *Med Phys*, 2017.
- [20] J. Milenkovic, O. Chambers, M. Marolt Music, and J. F. Tasic, "Automated breast-region segmentation in the axial breast mr images," *Comput Biol Med*, vol. 62, pp. 55–64, 2015.
- [21] D. M. Ufuk et al., "Using deep learning to segment breast and fibroglandular tissue in mri volumes," *Medical Physics*, vol. 44, no. 2, pp. 533–546, 2017.
- [22] D. L. Baggio, *GPGPU based image segmentation livewire algorithm implementation*. Thesis, 2007.
- [23] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.
- [24] E. Verburg et al., "Su-c-207b-04: Automated segmentation of pectoral muscle in mr images of dense breasts," *Med Phys*, vol. 43, no. 6, p. 3330, 2016.
- [25] J. Canny, "A computational approach to edge detection," *IEEE Trans Pattern Anal Mach Intell*, vol. 8, no. 6, pp. 679–98, 1986.
- [26] J. C. Bezdek, R. Ehrlich, and W. Full, "Fcm: The fuzzy c-means clustering algorithm," *Computers Geosciences*, vol. 10, no. 2, pp. 191–203, 1984.
- [27] E. W. Dijkstra, "A note on two problems in connexion with graphs," *Numerische Mathematik*, vol. 1, no. 1, pp. 269–271, 1959.
- [28] T. Sørensen, "A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons," *Biol. Skr.*, vol. 5, pp. 1–34, 1948.
- [29] A. M. Dinkla et al., "Mr-only brain radiation therapy: Dosimetric evaluation of synthetic cts generated by a dilated convolutional neural network," *International Journal of Radiation Oncology • Biology • Physics*, vol. 102, no. 4, pp. 801–812, 2018.

- [30] J. M. Wolterink, T. Leiner, M. A. Viergever, and I. Išgum, "Dilated convolutional neural networks for cardiovascular mr segmentation in congenital heart disease," in *Reconstruction, Segmentation, and Analysis of Medical Images* (M. A. Zuluaga et al., eds.), pp. 95–102, Springer International Publishing, 2017.
- [31] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *Corr*, vol. abs/1412.6980, 2014.
- [32] F. Milletari, N. Navab, and S. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 Fourth International Conference on 3D Vision (3DV)*, pp. 565–571, 2016.
- [33] D. P. Huttenlocher, G. A. Klanderman, and W. A. Rucklidge, "Comparing images using the hausdorff distance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 15, no. 9, pp. 850–863, 1993.
- [34] E. Morris et al., "Acr bi-rads® atlas, breast imaging reporting and data system," *Reston, VA: American College of Radiology*, pp. 56–71, 2013.
- [35] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* (N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, eds.), pp. 234–241, Springer International Publishing, 2015.
- [36] O. Bernard et al., "Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: Is the problem solved?," *IEEE Transactions on Medical Imaging*, vol. 37, no. 11, pp. 2514–2525, 2018.
- [37] P. Moeskops et al., "Automatic segmentation of mr brain images with a convolutional neural network," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1252–1261, 2016.
- [38] K. Kamnitsas et al., "Ensembles of multiple models and architectures for robust brain tumour segmentation," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries* (A. Crimi et al., eds.), pp. 450–462, Springer International Publishing, 2018.

2.7 Supplemental material 1: Robustness of the KBM

During robustness analyses of the KBM method, 81 different KBM methods were developed with a different combination of the weighting factors 2, 10 and 100 for parameters β , γ , δ and ϵ . The train datasets were segmented using all different models. All possible permutations of the factors and their resulting median DSC score on the train data are listed in the Table 2.4. In Figure 2.9 an overview of all DSC results is given in a box plot. Both table and figure show that permutation 41 is the best performing weighting factor combination. Permutation 41 was equal to the configuration used for the KBM as proposed in the manuscript.

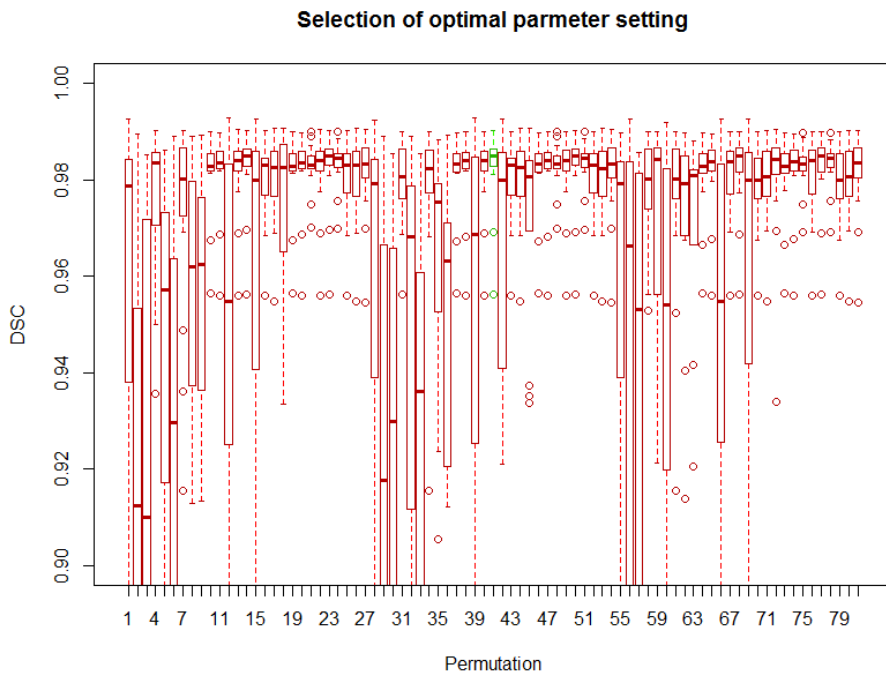


Figure 2.9: Boxplot of DSC results in the training data with 81 different KBM models resulting from all permutations. Permutation numbers correspond with the numbers used in Table 2.4. Permutation 41 (green) was equal to the configuration used for the KBM as proposed in the manuscript.

Table 2.4: All possible permutations of weighting factors 2, 10 and 100 for parameters β , γ , δ and ϵ with their corresponding DSC result. Each row starts with permutation number, followed by the used weighting factors and DSC score

#	β	γ	δ	ϵ	Median DSC result	#	β	γ	δ	ϵ	Median DSC result
1	2	2	100	10	0.979	42	10	10	2	2	0.980
2	2	2	100	100	0.912	43	100	10	2	10	0.983
3	2	2	100	2	0.910	44	100	10	2	100	0.982
4	10	2	100	10	0.984	45	100	10	2	2	0.981
5	10	2	100	100	0.957	46	2	100	2	10	0.983
6	10	2	100	2	0.930	47	2	100	2	100	0.984
7	100	2	100	10	0.980	48	2	100	2	2	0.983
8	100	2	100	100	0.962	49	10	100	2	10	0.984
9	100	2	100	2	0.962	50	10	100	2	100	0.985
10	2	10	100	10	0.983	51	10	100	2	2	0.984
11	2	10	100	100	0.984	52	100	100	2	10	0.983
12	2	10	100	2	0.955	53	100	100	2	100	0.982
13	10	10	100	10	0.984	54	100	100	2	2	0.983
14	10	10	100	100	0.985	55	2	2	10	10	0.979
15	10	10	100	2	0.980	56	2	2	10	100	0.966
16	100	10	100	10	0.983	57	2	2	10	2	0.953
17	100	10	100	100	0.983	58	10	2	10	10	0.980
18	100	10	100	2	0.982	59	10	2	10	100	0.984
19	2	100	100	10	0.983	60	10	2	10	2	0.954
20	2	100	100	100	0.984	61	100	2	10	10	0.980
21	2	100	100	2	0.983	62	100	2	10	100	0.979
22	10	100	100	10	0.984	63	100	2	10	2	0.981
23	10	100	100	100	0.985	64	2	10	10	10	0.983
24	10	100	100	2	0.985	65	2	10	10	100	0.984
25	100	100	100	10	0.983	66	2	10	10	2	0.955
26	100	100	100	100	0.983	67	10	10	10	10	0.984
27	100	100	100	2	0.983	68	10	10	10	100	0.985
28	2	2	2	10	0.979	69	10	10	10	2	0.980
29	2	2	2	100	0.918	70	100	10	10	10	0.980
30	2	2	2	2	0.930	71	100	10	10	100	0.981
31	10	2	2	10	0.981	72	100	10	10	2	0.984
32	10	2	2	100	0.968	73	2	100	10	10	0.983
33	10	2	2	2	0.936	74	2	100	10	100	0.984
34	100	2	2	10	0.982	75	2	100	10	2	0.983
35	100	2	2	100	0.975	76	10	100	10	10	0.984
36	100	2	2	2	0.963	77	10	100	10	100	0.985
37	2	10	2	10	0.983	78	10	100	10	2	0.984
38	2	10	2	100	0.984	79	100	100	10	10	0.980
39	2	10	2	2	0.969	80	100	100	10	100	0.981
40	10	10	2	10	0.984	81	100	100	10	2	0.984
41	10	10	2	100	0.985						

2.8 Supplemental material 2: Variations on DLM

In the manuscript we presented a 2D DCNN which segments the chest wall in transversal, coronal and sagittal directions and averages the results. For a performance comparison between different variations of the DLM we used the same train data and test data. In total, we trained 5 additional DLM methods. In 3 variation we used the same DCNN as presented in the manuscript, however we trained and tested it only on transversal, coronal or sagittal images. In the fourth variation, we used the same DNN structure to segmented transversal 2D slices without dilation rate. The last variation of the model was a 3D DCNN (Table 2.5). Results of all variation compared to the proposed DLM method in the manuscript are summarized in Table 2.6

Table 2.5: Configuration of the 3D DCNN.

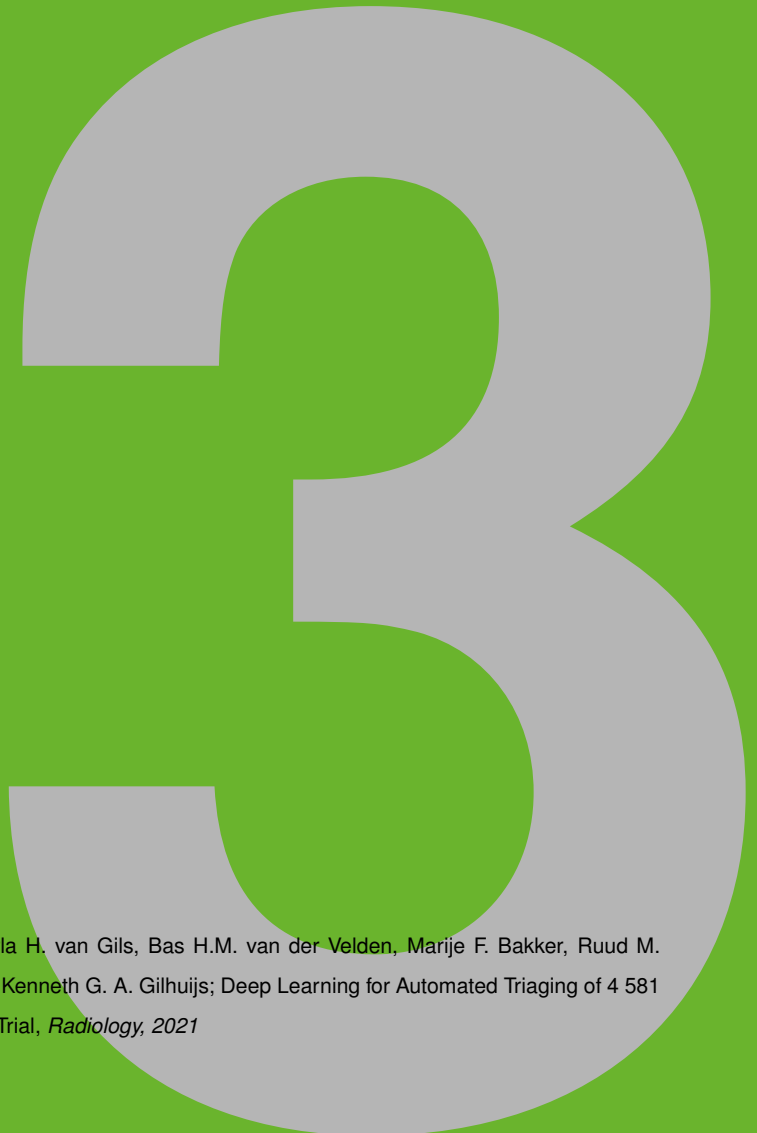
Layer	1	2	3	4	5	6	7	8	9	10
Convolution	3x3x3	3x3x3	3x3x3	3x3x1	3x3x1	3x3x1	3x3x1	3x3x1	1x1x1	1x1x1
Dilation	1x1x1	1x1x1	2x2x2	4x4x1	8x8x1	16x16x1	32x32x1	1x1x1	1x1x1	1x1x1
Field	3x3x3	5x5x5	9x9x9	17x17x9	33x33x9	65x65x9	129x129x9	131x131x9	131x131x9	131x131x9
Channels	32	32	32	32	32	32	32	32	192	3
Parameters	896	27680	27680	9248	9248	9248	9248	9344	6912	579

Table 2.6: Results of variations of the DLM method as proposed in the manuscript. All methods were compared to DLM, significant differences are marked by a *. The DLM transversal without dilation was also compared to DLM transversal, here significant differences are marked by a †

Proposed methods	Number of MRI datasets	DSC	p-value	FNF	p-value	FPF	p-value	HD95 (mm)	p-value	HD (mm)	p-value
DLM	79	0.982 (0.006)	-	0.003 (0.003)	-	0.030 (0.009)	-	2.58 (1.78)	-	12.81 (27.28)	-
DLM variation											
DLM Transversal	79	0.984(0.005)	0.18	0.003(0.003)	0.66	0.028(0.008)	0.31	3.75(6.16)	0.30	18.14(48.85)*	0.03
DLM Coronal	79	0.958(0.012)*	<0.01	0.043(0.020)*	<0.01	0.038(0.021)*	<0.01	11.1(7.10)*	<0.01	30.15(35.80)*	<0.01
DLM Sagittal	79	0.984(0.005)	0.38	0.006(0.004)*	<0.01	0.027(0.007)*	<0.01	4.59(3.98)*	<0.01	16.98(39.66)*	<0.01
DLM Transversal without dilation	79	0.971(0.012)*†	*<0.01	0.006(0.006)*†	*<0.01	0.048(0.021)*†	*<0.01	18.98(10.9)*†	*<0.01	47.82(27.02)*†	*<0.01
DLM 3D	79	0.984(0.004)	†<0.01	0.003(0.002)	†<0.01	0.028(0.006)	†<0.01	3.47(2.18)*	†<0.01	23.26(17.33)	†<0.01

Chapter 3

Deep Learning for Automated Triaging of 4 581 breast MRIs from the DENSE Trial



Based on: **Erik Verburg**, Carla H. van Gils, Bas H.M. van der Velden, Marije F. Bakker, Ruud M. Pijnappel, Wouter B. Veldhuis, Kenneth G. A. Gilhuijs; Deep Learning for Automated Triaging of 4 581 breast MRIs from the DENSE Trial, *Radiology*, 2021

3.1 Abstract

Background

MRI supplemental screening in women with extremely dense breasts has proved beneficial. Most MRIs show normal anatomical and physiological variation not requiring radiological review. Thus, ways to triage these normal MRIs to reduce radiologist workload are needed.

Purpose

To determine the feasibility of an automated triaging method using deep learning to dismiss the highest number of MRIs without lesions while still identifying malignant disease.

Materials and Methods

This secondary analysis of data from the DENSE trial evaluated breast MRIs from the first screening round of eight hospitals obtained between December 2011 and January 2016. A deep learning (DL) model was developed to distinguish between breasts with lesions and breasts without lesions. The model was trained to dismiss breasts with normal phenotypical variation and to triage lesions (BI-RADS 2-5) using eight-fold internal-external validation: trained on seven hospitals and tested on the eighth hospital, alternating such that each hospital once was an external test set. Performance was assessed using receiver-operating characteristic analysis. At 100% sensitivity for malignant disease, the fraction of examinations dismissed from radiological review was estimated.

Results

4581 MRI datasets of extremely dense breasts from 4581 women (mean age 54.3 years, IQR 51.5-59.8) were included. Of the 9162 breasts, 838 had at least one lesion (BI-RADS 2-5, of which 77 malignant) and 8324 had no lesions. At 100% sensitivity for malignant lesions, the DL model considered 90.7% (95%CI: 86.7%, 94.7%) of the MRIs with lesions to be non-normal and triaged them to radiological review. The DL model dismissed 39.7% (95%CI: 30.0%, 49.4%) of the MRIs without lesions. The DL model had an average AUC of 0.83 (95%CI: 0.80, 0.85) to discriminate between normal breast MRIs and MRIs with lesions.

Conclusion

Automated analysis of breast MRI in women with dense breasts dismissed nearly 40% of MRIs without lesions while not missing any malignant disease.

3.2 Introduction

Mammography is less sensitive in women with extremely dense breasts (American College of Radiology Breast Imaging Reporting and Data System [ACR BI-RADS] class D) than in women with fatty breasts[1–4]. Moreover, women with extremely dense breasts have a 3-6 times higher risk of developing breast cancer than women with almost entirely fatty breasts and a 2-fold higher risk than the average woman[5, 6]. Recently, a randomized controlled trial showed that supplemental screening with MRI in women between 50 and 75 years of age aids in detecting breast cancer at an earlier stage and significantly reduces the interval cancer rate between screening rounds[7].

In the Netherlands, approximately 8 percent of the screening participants have extremely dense breasts[8]. In a biennial screening program, nearly 82 000 women are eligible for MRI breast screening in the Netherlands alone[9]. Thus, routine MRI breast screening of women with extremely dense breasts on a population scale will be challenging. The workload for MRI operators and radiologists will increase substantially.

Several research directions aim at reducing the workload of MRI screening. These efforts include MRI protocols to reduce acquisition time without loss of sensitivity or specificity[10–14] and computer-aided diagnosis to reduce follow-up activities on benign findings[15, 16]. In an average-risk population, the vast majority (90.5%) of women screened with extremely dense breasts do not have findings that warrant further diagnostic workup[7]. Hence, it is of interest to automatically triage radiological review of breast MRI to women who have an above-average likelihood of harboring disease. If such an approach holds potential, health care resources could be prioritized to these higher-risk women, while reducing follow-up on benign findings.

This study is the first to investigate image-based triaging on multicenter screening data of women with extremely dense breasts at average risk. The aim of this research was to determine the feasibility of automated triaging using deep learning based on screening breast MRI to reduce the workload and to prioritize the work of breast MRI radiologists by dismissing the largest number of MRIs without lesions while still identifying all MRIs with malignant disease.

3.3 Materials and methods

3.3.1 Participants

In this secondary analysis of data from the first round of the prospective DENSE trial (ClinicalTrials.gov: NCT01315015), 4 783 MRI data sets were consecutively included from eight hospitals in the Netherlands between December 2011 and January 2016. When the MRI data sets were not present in full, they were excluded from this study. Data generated or analyzed during the study are available from the corresponding author by request. Participating women were recruited from the population-based breast cancer screening program in the Netherlands, which offers biennial mammography to women aged 50-75 years. The median age was 54 years (IQR 51–59 years)[7]. The trial was approved by the Dutch Minister of Health, Welfare and Sport (2011/19 WBO, The Hague, the Netherlands). According to the Dutch law on Population Studies, the study was hence waived from ethical review by the local IRB.

3.3.2 Imaging

The breast MRIs were acquired according to fixed imaging protocol described by Emaus et al[17]. In brief, the full multiparametric MRI protocol consisted of a high spatial resolution and high temporal resolution T1-weighted dynamic contrast-enhanced series, T2-weighted sequences and DWI sequences. Fat suppression was optional for both T1- and T2-weighted sequences.

This study focused only on the pre-contrast and first post-contrast images of the dynamic contrast-enhanced series at high spatial resolution (flip angle ranged between 10° and 20°, echo times between 1.7 ms and 2.4 ms and repetition time between 3.3 ms and 5.5 ms[17]) because these series are typically available in hospitals where breast DCE MRI is performed. All images were acquired using a 3-Tesla MRI unit; five hospitals used Philips MRI devices and three hospitals used Siemens MRI devices. The reconstructed voxel size depended on the MRI device (Table 3.1). All MRI examinations were performed in the axial plane with bilateral anatomic coverage, a field strength of 3.0 T, either a seven- or a 16-channel phased-array dedicated bilateral breast coil. Contrast agent was injected at a rate of 1 mL/sec to a total dose of 0.1-mmol of gadobutrol (Gadovist; Bayer AG) per kilogram of body weight.

Table 3.1: Overview of MRI devices and imaging properties of the dynamic contrast-enhanced series used during the first round of the DENSE trial

Medical Center	# Trial participants	MRI device	Reconstructed voxel size (mm ³)	Dimensions (voxels)	Fat suppression
1	1615	Philips Achieva	0.89x0.89x0.9	384x384x200	Yes
		Philips Ingenia	0.89x0.89x0.9	384x384x200	Yes
2	425	Siemens Magnetom Trio	0.80x0.80x1.0	448x448x176	No
		Siemens Skyra	0.80x0.80x1.0	448x448x176	No
		Siemens Prisma	0.80x0.80x1.0	448x448x176	No
3	244	Philips Achieva	0.89x0.89x0.9	384x384x200	Yes
4	500	Philips Ingenia	0.89x0.89x0.9	384x384x200	Yes
		Philips Achieva	0.89x0.89x0.9	384x384x200	Yes
5	316	Philips Ingenia CX	0.89x0.89x0.9	384x384x200	Yes
		Siemens Verio	0.85x0.85x1.0	448x448x176	No
7	489	Philips Ingenia	0.89x0.89x0.9	384x384x200	Yes
8	650	Siemens Skyra	0.80x0.80x1.0	448x448x160	No

3.3.3 Imaging Analysis in the DENSE trial

Imaging analysis in the DENSE trial is described elsewhere[7]. In brief, all MRIs were single read, and scored according to the BI-RADS MRI lexicon[18] by 16 trained breast MRI radiologists (among whom W.V.), whose experience ranged from 5 to 23 years in reading breast MRIs[7]. BI-RADS 3 lesions were double-read, recommended for repeat MRI after 6 months and subsequent biopsy on indication. BI-RADS 4 and BI-RADS 5 lesions were always indicated for biopsy. MRIs were ‘breasts with lesions’ (BI-RADS 2, 3, 4 or 5) or ‘breasts without lesions’. After negative biopsy, a participant returned to the screening workflow of the dense trial. Foci (<5 mm) were not taken into consideration, following consensus on their definition (i.e., focal enhancement too small to characterize any further[19]).

3.3.4 Deep learning

A method was developed to automatically establish whether an MRI of a breast contains a lesion, based on a deep learning (DL) model. The model was trained on left and right breasts separately, and combined into one result per MRI. The method follows three steps:

1. Image processing consisting of image cropping, registration and MIP creation (supplemental material 1: Image Preprocessing)

2. optimization of model architecture
3. internal-external validation[20, 21].

The network, source code and trained weights is fully available (https://github.com/Lab-Translational-Cancer-Imaging/AI_TriagingDENSE)

3.3.5 Model architecture optimization and internal-external validation

The model was developed by separating the DENSE data at hospital level, training and validating on seven hospitals, and using the remaining one as independent test hospital. This process was repeated eight times such that each hospital formed an independent test set (referred to as internal-external validation). The image data of the 7 hospitals were randomly separated in a train set (80%) and validation set (20%), the data of fold 8 (i.e., the independent test hospital) was the test set. Input image parameters and model parameters were part of the search for the optimal architecture of the neural network (supplemental material 2: CNN parameter optimization). The binary cross entropy loss function was used for optimization.

To properly train the neural network, the number of breasts that contain lesions was balanced against the number of breasts without lesions in the train set of seven hospitals, using random subsampling of the majority. For training, only one label 'yes' (i.e., lesion(s) present) or 'no' (i.e., no lesions present) was used as reference standard. The model with the smallest validation error was chosen as the final model. This model was applied to the test hospital. The model yields probability that a lesion is present in a breast. The maximum from left and right breast yields the probability that a bilateral MRI contains lesions. Results were stratified by MRI and lesion. The threshold was set at a value corresponding to 100% sensitivity for malignant lesions in the seven hospitals. Taking stochastic uncertainty at the extremes into account, the threshold was implemented at the 25th percentile of the probabilities of malignant disease observed in the train set minus the interquartile range, following the rationale proposed by Tukey[22].

3.3.6 Statistical Analysis

This study used all data acquired from the first screening round of the DENSE trial, sample size considerations for which are given in[7]. ROC curve analysis

was performed for each test hospital separately. All results are presented as the mean values across the eight observations, the confidence intervals for these means are derived from those eight observations.

Properties of the lesions to triage to radiological review were compared with those to be dismissed. Lesion size, volume and incidence of mass vs non-mass enhancement were compared using the student t-test and the distribution of BI-RADS score and Background Parenchymal Enhancement (BPE) were compared using one-way ANOVA. A p-value smaller than 0.05 was considered significant. In addition, the fraction of MRIs triaged without lesions was stratified by BPE and compared using one-way ANOVA. The index tests and statistical analyses in this secondary analysis of the DENSE data[7] were performed by EV and KG using R-Studio (Version 1.1.383, RStudio, Inc)

Because the model was trained on left and right breasts separately, we examined potential correlations of predicted lesion presence between left and right breasts (supplemental material 3: Correlation in predicted lesion presence between left and right breasts).

Because DL yields “black-box” results, an interpretation step was added to the network to visualize which image regions are responsible for triggering the MRI to triage. For this purpose, Deep Shapley Additive exPlanations (SHAP)[23] was used. In short, Deep SHAP visualizes the contribution of each voxel to the prediction result using a SHAP-map[24]: higher SHAP values correspond to higher probability of lesion presence. Using this SHAP-map, it was assessed whether the model prediction of lesion presence was based on plausible image regions.

3.4 Results

3.4.1 Participant Characteristics

In total, 4 581 of 4 783 MRI datasets (95.8%) of extremely dense breasts were included of 4 581 women aged between 50 and 75 years old (median age 54.3 years, IQR 51.5-59.8). Datasets were excluded (n=202) because the T1-weighted MRI data were not available in full due to incomplete acquisition or incomplete data transfer. Of the 9 162 extremely dense breasts (left and right), 838 breasts had at least one lesion (BI-RADS 2 or higher) and 8 324 breasts had no lesions. Seventy-seven malignant lesions were detected in 76 MRIs in the DENSE trial. Fifteen of these received BI-RADS-5 score, 57 BI-RADS 4, and 5 received an ini-

tial BI-RADS 3 score, upgraded to BI-RADS 4 on the 6-month-follow-up MRI (Table 3.2). Histopathology workup was based on core biopsies. Findings detected by the model but not found by radiologists in the DENSE trial, were considered unproven study findings, and did not trigger further patient workup.

Table 3.2: total number of single breasts in MRIs during training, stratified by lesion presence and BI-RADS score.

Medical Center	No lesion present	Lesion presence				Total
		BI-RADS 2	BI-RADS 3	BI-RADS 4	BI-RADS 5	
1	2 861	106	68 (4)	136 (23)	11 (8)	3 182
2	700	44	20 (1)	24 (7)	2 (1)	790
3	441	9	14	16 (2)	0	480
4	901	43	15	24 (3)	3 (2)	986
5	524	61	9	11 (2)	1 (1)	606
6	1 002	2	8	22 (4)	0	1 034
7	727	50	11	14 (5)	2 (2)	804
8	1 168	42	18	51 (10)*	1 (1)	1 280
Total	8 324	357	163 (5)	298 (56)*	20 (15)	9 162(76)*

the number of breasts in which a malignant lesion is present is shown in parentheses. *one breast in the dataset contained two malignant lesions.

3.4.2 Result of model architecture optimization

The best performing neural network architecture contained five blocks of two convolutional layers followed by one 2x2 max pooling layer (Figure 3.1). All convolutional layers were followed by a rectified linear unit (ReLU), while the last layer was a dense layer with softmax activation function. The total number of trainable parameters of the architecture was 1 345 922. A full overview of the optimized parameters is shown in supplemental material 2: CNN parameter optimization.

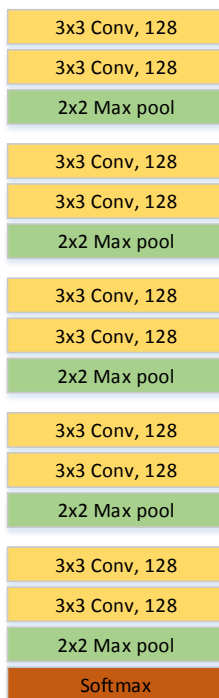


Figure 3.1: Schematic illustration of model architecture. Convolutional layers are shown by “Conv”, the final layer is the softmax activation function.

3.4.3 Performance of DL model

The average performance of the model across test hospitals was AUC=0.83 (95%CI 0.80-0.85) (Figure 3.2). At the threshold that detects all cancers 90.7% (95%CI: 86.7%, 94.7%) of the MRIs with lesions were considered to be non-normal (i.e., contained BI-RADS 2, 3, 4 or 5 lesions), and would be triaged to radiological review. Conversely, 39.7% (95%CI: 30.0%, 49.4%) of the MRIs without lesions would be dismissed. Because the model was trained to dismiss breasts with normal phenotypical variation and triage lesions, 88.4% (95%CI: 81.4%, 95.3%) of the BIRADS-2 MRIs exceeded the threshold to be considered normal and were triaged to radiological review. Among the largest groups here were fibroadenoma, indeterminate mass lesions and cysts.

BI-RADS 2 lesions were more likely to be dismissed (15.0% [95%CI: 6.0%, 23.9%]) than BI-RADS 4 lesions (8.8% [95%CI: 4.4%, 13.2%]) and BI-RADS 5 lesions (0%), $p=0.001$. We found no evidence of differences in lesion volume ($p=0.06$)

and lesion size ($p=0.49$). Non-mass lesions were more often dismissed than mass lesions ($p=0.01$). No evidence was found that BPE levels had impact on dismissal of lesions (Table 3.3). MRIs without lesions and minimal BPE were more often dismissed than MRIs without lesions and more severe BPE ($p<0.001$, Table 3.4).

The interpretable AI (SHAP) correctly visualized the image locations responsible for the prediction result (Figure 3.3), and correctly showed which breast was responsible for the triaging. SHAP values corresponding with low lesion probability were diffusely distributed. At higher probabilities of lesion presence, SHAP indicated that the model based its results on locations where lesions are present.

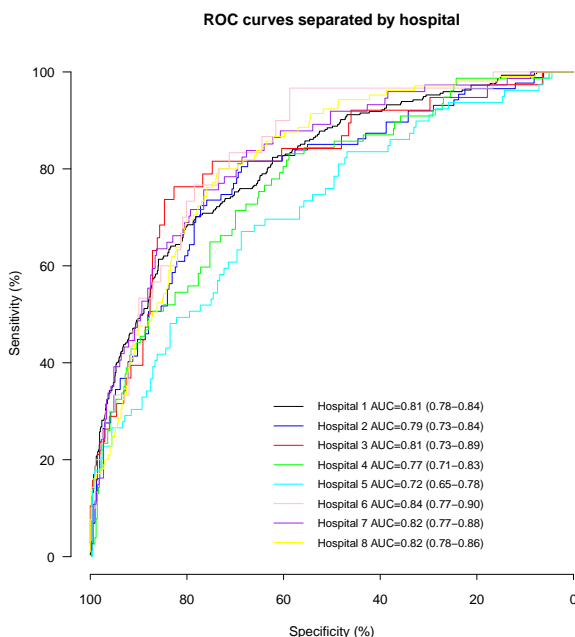


Figure 3.2: Receiver Operating Characteristics (ROC) curves of the eight hospitals. Each curve is the result of testing on one of the participating hospitals in the DENSE trial using internal-external validation. The curves show the sensitivity and the specificity of the method to distinguish between MRIs with lesions and MRIs without lesions. The area under the ROC curve (AUC) and 95% confidence interval are shown.

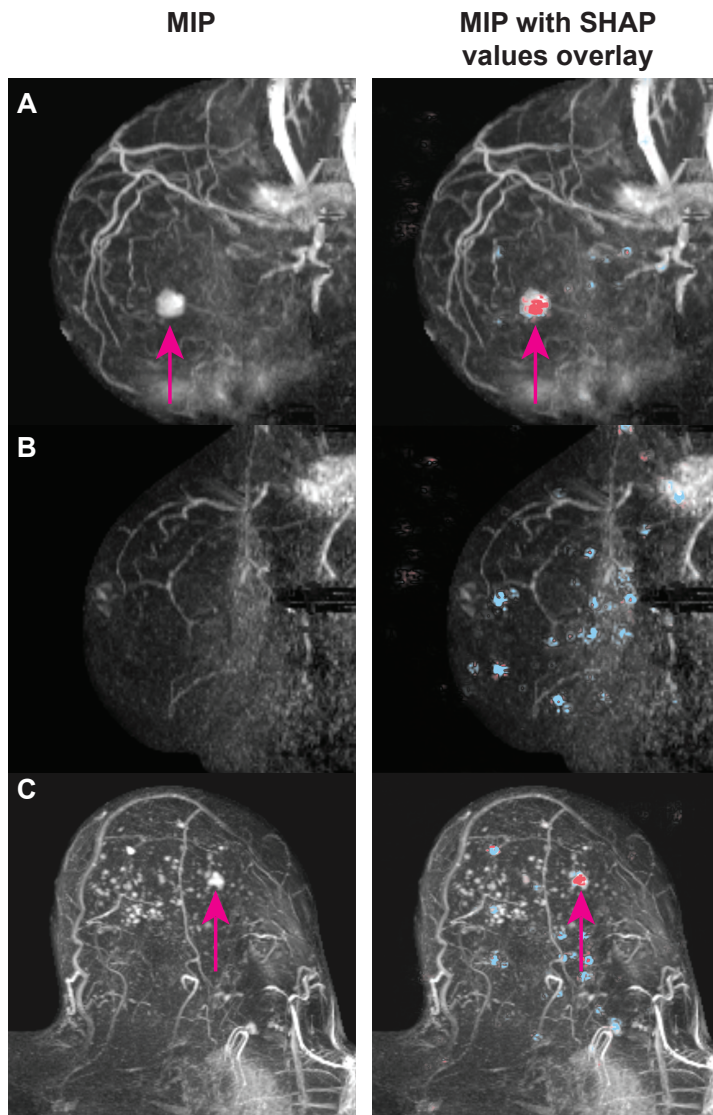


Figure 3.3: Examples of SHAP overlay images. On the left the maximum intensity image (MIP) and on the right the MIP with SHAP overlay. Positive SHAP values (red) show areas that contribute to a high probability of lesion presence, negative SHAP values (blue) show locations with reduced probability.

- A Sagittal MIP of contrast enhanced breast MRI of an invasive ductal carcinoma in a 57-year-old woman with BI-RADS 4 score. The DL yielded a probability of lesion presence of 90%. Positive SHAP values (red) are shown to coincide with the location of the lesion.
- B Sagittal MIP of contrast enhanced breast MRI of a breast without lesions in a 53-year-old woman with BI-RADS 1 score. The DL yielded a probability of lesion presence of 11%. Negative SHAP values (blue) are diffusely distributed in the breast region.
- C Transversal MIP of contrast enhanced breast MRI of a ductal carcinoma in situ in a 65-year-old woman with BI-RADS 4 score). The DL yielded a probability of lesion presence of 32%: the lowest probability value among all breasts with malignant disease in the current study. Positive SHAP values (red) are shown to coincide with the location of the lesion.

Table 3.3: Lesions to triage to radiological review using deep learning and those to be dismissed

Lesion or breast property	To triage to radiological review	To dismiss from radiological review	p-value
Volume (cm ³ , average and confidence interval)	0.59 (95%CI: 0.33, 0.86)	0.33 (95%CI: 0.19, 0.48)	p=0.06
Size (largest diameter, cm, average and confidence interval)	1.31 (95%CI: 1.19, 1.42)	1.22 (95%CI: 0.96, 1.48)	p=0.49
Mass /non mass			
Mass	88.7% (95%CI: 82.2%, 95.2%)	11.3% (95%CI: 4.7%, 17.8%)	p=0.01
Non-mass	73.6% (95%CI: 63.5%, 83.7%)	26.4% (95%CI: 13.3%, 36.5%)	
BI-RADS score			
BI-RADS 2	85.0% (95%CI: 76.1%, 94.0%)	15.0% (95%CI: 6.0%, 23.9%)	p=0.001
BI-RADS 3	88.3% (95%CI: 80.3%, 96.2%)	11.7% (95%CI: 3.8%, 19.7%)	
BI-RADS 4	91.2% (95%CI: 86.8%, 95.6%)	8.8% (95%CI: 4.4%, 13.2%)	
BI-RADS 5	100.0% (95%CI: 100%, 100%)	0% (95%CI: 0%, 0%)	
Background Parenchymal Enhancement			
Minimal	84.3% (95%CI: 74.1%, 94.4%)	15.7% (95%CI: 5.6%, 25.9%)	p=0.70
Mild	94.0% (95%CI: 90.4%, 97.6%)	6.0% (95%CI: 2.4%, 9.6%)	
Moderate	96.3 % (95%CI: 89.0%, 100%)	3.7% (95%CI: 0.0%, 11.0%)	
Marked	86.1 % (95%CI: 70.9%, 100%)	13.9% (95%CI: 0.0%, 29.1%)	

*numbers are shown at the 100% sensitivity operating threshold.

3.5 Discussion

An automated method was developed to triage breast MRIs of women with extremely dense breasts to radiological reading, aiming to dismiss normal MRIs (i.e., without BI-RADS 2,3,4 or 5 lesions). A deep learning (DL) model was trained to discriminate between breasts with lesions (in 785 MRIs) and breasts without le-

Table 3.4: MRIs without lesions to triage to radiological review using deep learning and those to be dismissed.

Background Parenchymal Enhancement	To triage to radiological review	To dismiss from radiological review	p-value
Minimal	39.1% (95%CI: 27.9%, 50.3%)	60.9% (95%CI: 49.7%, 72.1%)	p<0.001
Mild	67.3% (95%CI: 54.7%, 79.9%)	32.7% (95%CI: 20.1%, 45.3%)	
Moderate	77.4% (95%CI: 60.1%, 94.8%)	22.6% (95%CI: 5.2%, 39.9%)	
Marked	78.7% (95%CI: 60.3%, 97.1%)	21.3% (95%CI: 2.9%, 39.7%)	

sions (in 3 796 MRIs). MRIs were consecutively included from the first screening round of the DENSE trial. The model dismissed 39.7% (95%CI: 30.0%, 49.4%) normal breast MRIs without missing any malignant disease using internal-external validation in eight hospitals with various MRI devices. At this operating threshold, 90.7% (95%CI: 86.7%, 94.7%) of the MRIs with lesions would be triaged for confirmation by a radiologist. The methods performance, expressed by the area under the ROC curve, was 0.83 (95%CI: 0.80, 0.85). Accurate dismissal of normal MRIs without lesions depended somewhat on presence of BPE (i.e., a larger fraction of MRIs was dismissed when BPE was minimal). Although the model triaged a complete bilateral MRI even when only one of the breasts contained lesions, the method visualized the locations of the lesions that triggered the triaging. Only automated reduction of patient motion was applied as preprocessing. MRI quality was not curated in any other way. Hence, the data contained typical artifacts that occur in daily clinical practice, and the performance was deemed representative for typical screening MRIs. Multiple other publications reported findings on detection and classification of breast lesions on MRI[25–29]. These studies are difficult to compare to the current study because they do not report the fraction of correctly identified breasts that do not contain lesions. The most comparable approach to our study was used by Gubern-Mérida et al.[25]. This study did not aim, however, at detection of normal breasts, and reported seven false-positive findings on average in breasts without lesions.

The explainable DL (using SHAP) provided fast interpretation of why the system reached the decision to triage an MRI. The challenge here is to increase the confidence of the system in the “gray zone” of normal anatomical and physiological

variation in dense breasts that can also be subtle signs of underlying malignant disease. The system was set to the safest operating threshold to prevent false-negative dismissals. The setting may be optimized over time as the volume of training data increases.

The best performance was achieved using three MIP directions (transversal, sagittal, and coronal), indicating that useful information exists in the combination of these planes. More complexity in the model, i.e., more convolutional layers, additional dense layers, or additional input channels, did not improve the performance. Our study has limitations. The available sample size for training varied in each fold during internal-external validation. The largest fraction of data was obtained in hospital 1. Hence, the smallest fraction of training data was available for this hospital. To overcome this, it is possible to sub split the data of hospital one in multiple folds. The disadvantage of sub splitting during training is that not all folds are from different hospitals. It was therefore not our primary method of analysis. In addition, the results from individual folds in the internal-external validation did not indicate reduced model performance for the first hospital.

Our results are based on data obtained during the first round of the DENSE screening trial. The number of detected malignant lesions was smaller in subsequent biennially screening rounds of DENSE compared with the first round[30]. Cancers were smaller and less developed in the subsequent rounds. Therefore, we plan to further validate the performance of the model on data of subsequent rounds. The model should also be validated in consecutive datasets of varying quality from more hospitals, for example to assess the minimal quality requirements, and to further document the robustness and usability in clinical practice. Such information will be essential for certification of the AI for clinical use in the future.

Future research could focus on post-hoc analysis of triaged lesions into benign or malignant classes using computer-aided diagnosis. In addition, prospective controlled trials and registration studies will have to demonstrate that the AI is at least as effective at dismissing normal MRIs as the expert breast radiologist.

In conclusion, a deep learning model was developed that identified breast MRIs without cancer with high certainty. Using internal-external validation, the method identified nearly 40% of the MRIs with normal anatomical and physiological variation in women with extremely dense breasts without missing any malignant disease.

References

- [1] R. E. Bird, T. W. Wallace, and B. C. Yankaskas, "Analysis of cancers missed at screening mammography," *Radiology*, vol. 184, no. 3, pp. 613–617, 1992.
- [2] N. F. Boyd et al., "Mammographic density and the risk and detection of breast cancer," *N Engl J Med*, vol. 356, no. 3, pp. 227–36, 2007.
- [3] K. Kerlikowske, "The mammogram that cried wolfe," *New England Journal of Medicine*, vol. 356, no. 3, pp. 297–300, 2007.
- [4] J. O. Wanders et al., "Volumetric breast density affects performance of digital screening mammography," *Breast Cancer Res Treat*, vol. 162, no. 1, pp. 95–103, 2017.
- [5] E. R. Price et al., "The california breast density information group: a collaborative response to the issues of breast density, breast cancer risk, and breast density notification legislation," *Radiology*, vol. 269, no. 3, pp. 887–92, 2013.
- [6] V. A. McCormack and I. dos Santos Silva, "Breast density and parenchymal patterns as markers of breast cancer risk: a meta-analysis," *Cancer Epidemiol Biomarkers Prev*, vol. 15, no. 6, pp. 1159–69, 2006.
- [7] M. F. Bakker et al., "Supplemental mri screening for women with extremely dense breast tissue," *New England Journal of Medicine*, vol. 381, no. 22, pp. 2091–2102, 2019.
- [8] D. van der Waal et al., "Geographic variation in volumetric breast density between screening regions in the netherlands," *European Radiology*, vol. 25, no. 11, pp. 3328–3337, 2015.
- [9] IKNL, "Monitor bevolkingsonderzoek borstkanker," 2018.
- [10] C. K. Kuhl et al., "Abbreviated breast magnetic resonance imaging (mri): first postcontrast subtracted images and maximum-intensity projection—a novel approach to breast cancer screening with mri," *J Clin Oncol*, vol. 32, p. 2304, 2014.
- [11] S. C. Harvey et al., "An abbreviated protocol for high-risk screening breast mri saves time and resources," *J Am Coll Radiol*, vol. 13, p. 374, 2016.
- [12] B. Panigrahi et al., "An abbreviated protocol for high-risk screening breast magnetic resonance imaging: Impact on performance metrics and bi-rads assessment," *Academic Radiology*, vol. 24, no. 9, pp. 1132–1138, 2017.
- [13] Y. Machida et al., "Feasibility and potential limitations of abbreviated breast mri: an observer study using an enriched cohort," *Breast Cancer*, vol. 24, no. 3, pp. 411–419, 2017.
- [14] J. C. van Zelst et al., "Multireader study on the diagnostic accuracy of ultrafast breast magnetic resonance imaging for breast cancer screening," *Investigative radiology*, vol. 53, no. 10, pp. 579–586, 2018.
- [15] M. U. Dalmiş et al., "Artificial intelligence–based classification of breast lesions imaged with a multiparametric breast mri protocol with ultrafast dce-mri, t2, and dwi," *Investigative Radiology*, vol. Publish Ahead of Print, 2019.

- [16] E. Verburg et al., "Computer-aided diagnosis in multiparametric magnetic resonance imaging screening of women with extremely dense breasts to reduce false-positive diagnoses," *Invest Radiol*, vol. 55, no. 7, pp. 438–444, 2020.
- [17] M. J. Emaus et al., "Mr imaging as an additional screening modality for the detection of breast cancer in women aged 50-75 years with extremely dense breasts: The dense trial study design," *Radiology*, vol. 277, no. 2, pp. 527–37, 2015.
- [18] E. Morris et al., "Acr bi-rads® atlas, breast imaging reporting and data system," Reston, VA: American College of Radiology, pp. 56–71, 2013.
- [19] R. M. Mann, N. Cho, and L. Moy, "Breast mri: State of the art," *Radiology*, vol. 292, no. 3, pp. 520–536, 2019.
- [20] P. Royston, M. K. B. Parmar, and R. Sylvester, "Construction and validation of a prognostic model across several studies, with an application in superficial bladder cancer," *Statistics in Medicine*, vol. 23, no. 6, pp. 907–926, 2004.
- [21] E. W. Steyerberg et al., "Predicting outcome after traumatic brain injury: development and international validation of prognostic scores based on admission characteristics," *PLoS medicine*, vol. 5, no. 8, p. e165, 2008.
- [22] J. W. Tukey, *Exploratory Data Analysis*. Addison-Wesley, 1977.
- [23] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in neural information processing systems*, pp. 4765–4774, 2017.
- [24] B. H. M. van der Velden et al., "Interpretable deep learning regression for breast density estimation on MRI," in *Medical Imaging 2020: Computer-Aided Diagnosis* (H. K. Hahn and M. A. Mazurowski, eds.), vol. 11314, pp. 253 – 258, International Society for Optics and Photonics, SPIE, 2020.
- [25] A. Gubern-Mérida et al., "Automated localization of breast cancer in dce-mri," *Medical Image Analysis*, vol. 20, no. 1, pp. 265–274, 2015.
- [26] M. U. Dalmış et al., "Fully automated detection of breast cancer in screening mri using convolutional neural networks," *Journal of medical imaging (Bellingham, Wash.)*, vol. 5, no. 1, pp. 014502–014502, 2018.
- [27] A. Vignati et al., "Performance of a fully automatic lesion detection system for breast dce-mri," *Journal of Magnetic Resonance Imaging*, vol. 34, no. 6, pp. 1341–1351, 2011.
- [28] D. M. Renz et al., "Detection and classification of contrast-enhancing masses by a fully automatic computer-assisted diagnosis system for breast mri," *Journal of Magnetic Resonance Imaging*, vol. 35, no. 5, pp. 1077–1088, 2012.
- [29] Y.-C. Chang et al., "Computerized breast lesions detection using kinetic and morphologic analysis for dynamic contrast-enhanced mri," *Magnetic resonance imaging*, vol. 32, no. 5, pp. 514–522, 2014.

- [30] M. Bakker et al., "Mri in addition to mammography screening in women with extremely dense breasts: Primary outcome of the randomized dense trial," *Radiological Society of North America 2019 Scientific Assembly and Annual Meeting*, 2019.
- [31] E. Verburg et al., "Knowledge-based and deep learning-based automated chest wall segmentation in magnetic resonance images of extremely dense breasts," *Medical physics*, 2019.
- [32] S. Klein et al., "Elastix: A toolbox for intensity-based medical image registration," *IEEE Transactions on Medical Imaging*, vol. 29, no. 1, pp. 196–205, 2010.
- [33] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
- [34] J. S. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl, "Algorithms for hyper-parameter optimization," in *Advances in neural information processing systems*, pp. 2546–2554, 2011.
- [35] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *Corr*, vol. abs/1412.6980, 2014.

3.6 Supplemental Material 1: Image pre-processing

3.6.1 Image preprocessing

The image preprocessing consisted of three steps:

1. Fully-automated cropping of images to the breast region using dynamic programming[31].
2. Image registration to correct for motion, which uses a non-rigid B-spline transformation in a multiresolution scheme[25, 32].
3. Creation of 2D maximum intensity projections (MIP) from subtracted datasets of the left and right breast separately in three directions (coronal, sagittal, and transversal). The size of the cropped images depends on individual breast size. Thus, the images were extended by zero padding to achieve uniform dimensions of 256 x 256 pixels. No downscaling was applied.

The results of the three automated steps were visually checked by a technical physician (E.V., 6 years of experience) under supervision of a breast MRI radiologist (W.V., 11 years of experience). No manual corrections were required.

3.7 Supplemental Material 2: CNN parameter optimization

Convolutional neural networks VGG16 and VGG19 DL models were used as basis for model development[33]. The model architecture was optimized using the Neural Network Intelligence tool (NNI, Version 1.0, Microsoft) in combination with the Tree-structured Parzen Estimator (TPE) tuner[34] with a maximum of 500 trials. Experiments were performed using an NVIDIA GeForce RTX2080 GPU. The trainset of the first fold was used by the NNI to determine the optimal network architecture. The same architecture was used in all subsequent folds.

VGG networks consist of blocks of convolutional layers followed by a max pooling layer. Both the number of blocks and the number of convolutional layers were optimized. Also, the size of the layer, the convolutional kernel size, the number of dense layers in front of the last layer, the size of the dense layer, dropout rate, optimizer and loss function were optimized. The input was optimized for direction of the MIP and the number of directions. More specifically, during optimization, input varied from single MIP image only (in coronal, transversal, or sagittal direction), all combinations of two MIP images, and three MIP images from the same breast. When multiple MIPs of a breast were used as input, the multiple outputs were averaged per breast. Image normalization and augmentation settings were also optimized. Experiments were performed using an NVIDIA GeForce RTX2080 GPU. The trainset of the first fold was used by the NNI to determine the optimal network architecture. The same architecture was used in all subsequent folds.

The best performing neural-network architecture contained five blocks of two convolutional layers followed by one 2x2 max pooling layer. Each convolutional layer contained 128 filters using 3x3 convolutional kernels. All convolutional layers were followed by a rectified linear unit (ReLU), while the last layer was a softmax activation function. The mini batch size was 2, the maximal number of epochs was 20, the Adam optimizer[35] was used and the learning rate was optimized at $8.12 \cdot 10^{-5}$. No dense layers were added and no kernel dilation was applied. The kernel stride was 1 in all cases (Table 3.5). The total number of trainable parameters of the network was 1 345 922.

The optimal input to the model was found to contain 2D orthogonal maximum intensity projection (MIP) images of contrast uptake in both breasts (left and right) separately. Best results were obtained when the model was trained on MIP im-

ages in all three directions (i.e., transversal, sagittal, and coronal)[31], indicating that there is useful information in a combination of these planes to incorporate 3D information. Contrast uptake images were obtained by subtraction of the pre-contrast DCE MRI series from the first post contrast series. Intensities of the MIP images were normalized to zero mean unit variance. The results of the model were averaged in the three directions to obtain one probability of lesion presence. The number of breasts as part of training group, validation group and test group varied for each fold of the internal external validation (Table 3.6)

Table 3.5: Overview of the varied parameters and optimal parameters of the best performing architecture and input image settings as selected by NNI.

Parameter	Options	NNI optimized
Model architecture parameters		
Number of blocks of convolutional layers followed by a max pooling layer	2, 3, 4, 5	5
Number of convolutional layers in a block	2, 3, 4, 5	2
Layer size	64, 128, increasing over blocks*	128
Kernel size	3, 5	3
Number of dense layers	0, 1, 2	0
Size of dense layer	16,32,64	NA
Dropout rate in dense layer	Between 0 and 0.5	NA
Optimizer	Adam, AdaMax, AdaGrad, AdaDelta, Stochastic gradient descent, Stochastic gradient descent with Nesterov momentum	Adam
Loss function	Binary crossentropy, Tversky with alpha varying between 0.5 and 0.9	Binary crossentropy
Mini batch size	2,4,8	2
Learning rate	Between 0.00001 and 0.001	0.0000812
Model input parameters		
MIP directions	Sagittal, transversal, coronal, Sagittal + transversal, sagittal + coronal, transversal + coronal, sagittal + transversal + coronal	sagittal + transversal + coronal
Normalisation	None, zero-mean unit variance	Zero-mean unit variance
Augmentation	None, Flip left-right, flip left-right up-down, flip left right up-down and translation, flip left right up-down and translation and rotation	None

*Increasing over blocks means that the first block of convolution layers followed by a max pooling layer has a layer size of 32, the second block has a layer size of 64, the third has a layer size of 128, the fourth has layer size of 256, and the fifth has a layer size of 512. Which is identical to the VGG architecture.

Table 3.6: Number of breasts part of training group, validation group and test group during the internal external validation

Internal external validation fold number	Training (80%)		Validation (20%)		Test	
	Number of breasts with lesions	Number of breasts without lesions	Number of breasts with lesions	Number of breasts without lesions	Number of breasts with lesions	Number of breasts without lesions
1	414	414	103	103	321	2 863
2	598	598	150	150	90	700
3	639	639	160	160	39	441
4	602	602	151	151	85	901
5	605	605	151	151	82	524
6	645	645	161	161	32	1 002
7	609	609	152	152	77	727
8	581	581	145	145	112	1 168

3.8 Supplemental Material 3: Correlation in predicted lesion presence between left and right breasts

Because the model was trained on left and right breasts separately, we verified that the DL output for bilateral MRIs is not biased towards similar results in left and right breast by correlation through participant. For this purpose, the correlation in predicted lesion presence between left and right breasts of the same participants was calculated using the Spearman correlation.

We did not find evidence of correlations in predicted lesion presence between left and right breasts in bilateral MRIs. The correlation in lesion probability left and right was moderate when lesions were absent (Spearman's correlation coefficient $r = 0.56$, $p < 0.01$). The correlation in MRIs with bilateral lesions was also moderate (Spearman's correlation coefficient, $r = 0.69$, $p < 0.01$). We found no evidence of correlation in women with unilateral lesions (Spearman's correlation coefficient $r = 0.06$, $p < 0.01$).

Chapter 4

Computer-aided diagnosis in multi-parametric MRI screening of women with extremely dense breasts to reduce false positive diagnoses



Based on: **Erik Verburg**, Carla H. van Gils, Marije F. Bakker, Max A. Viergever, Ruud M. Pijnappel, Wouter B. Veldhuis, Kenneth G. A. Gilhuijs; Computer-aided diagnosis in multi-parametric MRI screening of women with extremely dense breasts to reduce false positive diagnoses, *Investigative Radiology*, 2020, Volume 55 (438-444)

4.1 Abstract

Objective

To reduce the number of false positive diagnoses in the screening of women with extremely dense breasts using Magnetic Resonance Imaging (MRI), we aimed to predict which BI-RADS-3 and BI-RADS-4 lesions are benign. For this purpose, we use computer-aided diagnosis (CAD) based on multi-parametric assessment.

Materials and Methods

Consecutive data were used from the first screening round of the Dense Tissue and Early Breast Neoplasm Screening (DENSE) trial. In this trial, asymptomatic women with a negative screening mammography and extremely dense breasts were screened using multi-parametric MRI. In total, 4783 women, aged 50-75 years, enrolled and were screened in 8 participating hospitals between December 2011 and January 2016. In total 525 lesions in 454 women were given a BI-RADS 3 (n=202), 4 (n=304) or 5 score (n= 19). Of these lesions, 444 were benign and 81 were malignant on histologic examination. The MRI protocol consisted of five different MRI sequences: T1-weighted imaging without fat suppression, diffusion-weighted imaging, T1-weighted contrast-enhanced images at high spatial resolution, T1-weighted contrast-enhanced images at high temporal resolution, and T2-weighted imaging. A machine learning method was developed to predict, without deterioration of sensitivity, which of the BI-RADS 3 and BI-RADS 4 scored lesions are actually benign and could be prevented from being recalled. BI-RADS 5 lesions were only used for training, because the gain in preventing false-positive diagnoses is expected to be low in this group. The CAD consists of two stages: feature extraction and lesion classification. Two groups of features were extracted; the first based on all multi-parametric sequences, the second based only on sequences that are typically used in abbreviated MRI protocols. In the first group, 49 features were used as candidate predictors: 46 were automatically calculated from the MR images, supplemented with 3 clinical features (age, BMI and BI-RADS score). In the second group, 36 image features and the same 3 clinical features were used. Each group was considered separately in a machine-learning model to differentiate between benign and malignant lesions. We developed a Ridge regression model using 10-fold cross validation. Performance of the models was analyzed using an accuracy measure curve and

receiver-operating characteristic analysis.

Results

Of the total number of BI-RADS 3 and BI-RADS 4 lesions referred to additional MRI or biopsy, 425/487 (87.3%) were false positive. The full multi-parametric model classified 176 (41.5%) and the abbreviated-protocol model classified 111 (26.2%) of the 425 false-positive BI-RADS 3 and BI-RADS 4 scored lesions as benign without missing a malignant lesion. If the full multi-parametric CAD had been used to aid in referral, recall for biopsy or repeat MRI could have been reduced from 425/487 (87.3%) to 311 / 487 (63.9%) lesions. For the abbreviated protocol, it could have been 376 / 487 (77.2%)

Conclusion

Dedicated multi-parametric CAD of breast MRI for BI-RADS 3 and 4 lesions in screening of women with extremely dense breasts has the potential to reduce false positive diagnoses and consequently to reduce the number of biopsies without missing cancers.

4.2 Introduction

Women with extremely dense breasts (Breast Imaging Reporting and Data System (BI-RADS) class D), i.e., breasts containing a large amount of fibroglandular tissue, have a 3-6 times higher risk of developing breast cancer than women with very fatty breasts. Moreover, these cancers are harder to detect on mammography due to the low contrast between fibroglandular tissue and tumor tissue and overlapping tissue[1]. Consequently, additional screening modalities, such as magnetic resonance imaging (MRI) have been proposed.

MRI is known to be a sensitive method to detect lesions. Several studies showed that additional MRI screening increases the number of detected malignancies[2–5]. However, MRI is also associated with lower specificity than mammography[5, 6]. Moreover, MRI is more costly and time-consuming than mammography. The effectiveness of additional MRI for the screening of women with extremely dense breasts is the main research aim of the Dense Tissue and Early Breast Neoplasm Screening (DENSE) trial in the Netherlands. Within the framework of this randomized controlled trial, 4,783 women with extremely dense breasts have been screened using additional MRI after a negative screening mammography[7, 8].

As anticipated, additional breast cancers were detected; in the first round of this trial the cancer detection yield with MRI after negative mammography was 79 in 4,783 women, or 16.5/1000 screens[8]. Subsequently, women in the MRI-arm experienced a significantly lower number of interval cancers than those in the control arm[8]. However, in total, 454 women (9.5%) were referred for additional diagnostics after MRI. BI-RADS 3 lesions led to recommendation for repeat MRI screening after 6 months and subsequent biopsy on indication. For women with BI-RADS 4 or BI-RADS 5 lesions, biopsy was indicated.

For women with BI-RADS 4 or BI-RADS 5 lesions, biopsy was indicated. As expected, the percentage of benign findings in BI-RADS 3 and BI-RADS 4 scored women was high. No malignant lesion was present in 97% of BI-RADS-3 scored women (146 out of 150) and 79% of the BI-RADS-4 scored women (226 out of 286). In BI-RADS 5 scored women 17% (3 out of 18) had no malignant lesions. Especially in BI-RADS 3 and BI-RADS 4 scored women, increased specificity would lead to reduced follow up activities.

Reports on different, heterogeneous populations of women showed potential for computer-aided diagnosis (CAD) to improve the specificity of breast MRI[9–11]. To the best of our knowledge, no studies focused explicitly on a consecutively in-

cluded screening population of asymptomatic women with extremely dense breasts and average risk.

Typically, CAD for breast MRI is based on dynamic contrast-enhanced T1-weighted images[12], but combination with other sequences have been used as well (i.e., multi-parametric MR imaging)[11, 13, 14]. In particular, high-temporal resolution dynamic contrast-enhanced series (fast-DCE)[13], diffusion-weighted imaging (DWI)[15] and T2-weighted imaging[16] have shown complementary value to discriminate between malignant and benign lesions. To reduce the number of false positive diagnoses in the MRI screening of women with extremely dense breasts, the aim of this study is to predict which BI-RADS 3 and BI-RADS 4 lesions are benign using multi-parametric CAD.

4.3 Materials and Methods

4.3.1 Study population

Clinical data and MRIs were obtained during the first round of the DENSE trial. The DENSE trial has been described in detail elsewhere[7]. In short, this multi-center randomized controlled trial investigates the additional value of MRI screening in Dutch women with extremely dense breasts (i.e., BI-RADS D and normal mammography). Written informed consent was obtained from all women before MRI screening. The trial was approved by the Dutch Minister of Health, Welfare and Sport (2011/19 WBO, The Hague, the Netherlands). In this study, all image datasets were acquired between 22 December 2011 and 22 January 2016. All women with lesions that were scored as BI-RADS 3, 4 or 5 on MRI were included in the analysis described here. Some women with an indication for biopsy (31 of 331) did not undergo a biopsy because, for example, the lesion was not/no longer visible on additional imaging, the biopsy was technically not possible (in which case short-term follow-up imaging was applied), or the lesion was known to be benign from the patient records from another hospital[8]. The median age of the participants was 54 years (range 49 to 75 years)

4.3.2 Study population

All breast MR images were acquired according to a fixed imaging protocol as described by Emaus et al[7]. In summary, the examinations were performed with a

3.0 T (Achieva or Ingenia) system from Philips or a 3.0 T (Trio, Verio or Skyra) system from Siemens using a dedicated phased-array bilateral breast coil. Images were acquired in axial planes. The MR imaging protocol consisted of DWI, T1-weighted imaging without fat suppression, DCE-MR, and an optional T2-weighted sequence. DCE-MR consisted of a high-spatial-resolution pre-contrast image, followed by a high-temporal-resolution series after contrast agent injection, followed by 4 or 5 high-spatial-resolution images. Fat suppression was optional during DCE-MR acquisition. The high-temporal-resolution series were acquired in 3.9 to 5.1 s intervals and consisted of 15 to 19 post-contrast acquisitions. Contrast agent was injected at a rate of 1 mL/sec to a total dose of 0.1 mmol of macrocyclic GBCA gadobutrol (Gadovist®, Bayer AG, Leverkusen, Germany) per kilogram of body weight. DWI was acquired with a minimum of two b-values and a maximal b-value of at least 800[7].

4.3.3 Methods

A CAD model was developed and tested to predict whether lesions on MRI in women with extremely dense breasts are benign or malignant. The first stage of the CAD workflow was image processing (section Image processing) followed by automated calculation of features from all BI-RADS 3, 4 and 5 lesions (section 3.3.2). The features were used to train and validate the model using cross validation (section 3.3.3). These steps were repeated for a subset of images typically available in abbreviated MRI protocols[9, 17], i.e. T2-weighted imaging, DWI and DCE-MR imaging consisting of high-temporal-resolution series, and one pre- and one post-contrast image with a high spatial resolution.

Image processing

Seven consecutive image processing steps were performed: (1) Image registration of DCE-MR series, (2) lesion segmentation, (3) DCE-MR image normalization, (4) aorta segmentation, (5) chest wall segmentation and extraction of pectoral muscle intensity, (6) calculation of apparent diffusion coefficient (ADC) and (7) registration of lesion mask to ADC map.

1 Image registration of DCE-MR images: All post-contrast DCE-MR images with high spatial resolution were registered to their pre-contrast counterparts using a non-rigid B-spline transformation in a multi-resolution scheme[18].

2 Lesion segmentation: The semi-automated segmentation method proposed by Alderliesten et al.[19] was used for lesion segmentation of mass lesions as well as non-mass lesions. Lesions were detected by breast radiologists associated with the DENSE trial, and whose experience ranged from 5 to 23 years[8]. A seed point was manually placed at or near the lesion by Technical Physician (EV). Subsequently, constrained volume growing was performed in the DCE-MR series. This step resulted in a segmented lesion volume in 3D. Segmentations were reviewed by a trained breast radiologist (WBV) and corrected when necessary by adding or replacing seed points.

3 DCE-MR image normalization: Although all images were acquired according to the screening protocol, some variations were present in the settings of the different MRI devices used, mainly flip angle and repetition time. Changes in intensity due to the inflow of contrast agent depend on these settings, and may therefore differ between hospitals. Hence, we normalized intensities by calculating the signals that would be acquired at a standard flip angle and repetition time[20, 21]. All DCE images with high spatial resolution were harmonized to a standard flip angle of 10° and a standard repetition time of 3.78 ms. All DCE images with high temporal resolution were harmonized to a standard flip angle of 10° and a standard repetition time of 2.17 ms.

4 Aorta segmentation: Contrast uptake speed in the lesion is related to contrast uptake in the descending aorta of the subject. Accordingly, the descending aorta was segmented in the DCE-MR images with high temporal resolution. The aorta was located on the basis of its tubular shape. We used the Hough transform to detect one circle with a diameter between 1 and 5 cm in each transversal slice in the last post-contrast series of the fast acquisition. A linear Hough transform was used to detect the main axis of the descending aorta. All found circles centered at the detected main axis were defined as the contours of the descending aorta.

5 Chest wall segmentation and extraction of pectoral muscle intensity: T2-weighted image intensities were normalized to the intensity of the pectoral muscle(16). First, the pectoral muscle was automatically detected in the T1-weighted images without fat suppression using dynamic programming[22]. Next, the detected chest wall was re-sampled to the dimensions of the T2-weighted image. The median

intensities ($MI(d)$) of the voxels located at nearest distance $d=0, 1, 2 \dots n$ mm medial from the chest wall in the T2-weighted image were calculated. The pectoral muscle intensity was defined as $MI(d)$ where the first local minimum or local maximum was present in $MI(d)$.

6 Calculation of apparent diffusion coefficient: ADC values for all voxels in the DWI images of each subject were calculated using a linear least squares estimator based on QR decomposition. ADC values were computed using the non-weighted image (b-value = 0) in combination with all individual diffusion-weighted images (b-value > 0), which resulted in one ADC map per subject.

7 Registration of lesion mask to ADC map: DWI images are susceptible to artifacts such as geometric distortions due to magnetic susceptibility differences[23]. This geometric distortion was corrected by registration as follows: First the T2-weighted images were registered to the corresponding DWI image with b-value 0 or 50. Non-rigid B-spline transformation in a multi resolution scheme was used[18]. Because the lesion mask is inherently aligned with the lesion on T2-weighted imaging, the transformation from T2 to DWI was applied to the lesion mask in order to align the mask with the lesion on ADC. Manual adjustment was applied by a Technical Physician (EV) when necessary.

Image registration was performed using Elastix (version: 4.7)[24], MeVisLab (version 3.0, MeVis medical Solution AG, Bremen, Germany) in combination with Python (version 2.7, Python Software Foundation) with packages 'numpy' (v1.15.1) and 'scipy' (v1.1.0) was used for lesion segmentation, image normalization and aorta segmentation. Chest wall segmentation and ADC map calculation were performed using MATLAB (v R2017a; Mathworks, Natick, MA).

Feature extraction

In total, 49 features were calculated and used to train the CAD model. All 46 MRI-based features were obtained automatically. Twenty-two features describing morphology and contrast dynamics of the lesion were computed from the high-spatial DCE images[12, 25].

Six contrast uptake features were computed from the fast-DCE images using a method based on the work of Dalmiş et al[13]. Here, time-related features were expressed relative to the start of contrast uptake in the detected descending aorta.

Nine ADC features and nine T2 intensity features were computed (see Table, Supplemental Material 1: Description of image features). In addition to image features, three clinical features were considered in the model: BI-RADS score (3, 4 or 5), age, and body mass index (BMI).

Missing features (506 out of 24 794) caused by missing images (n=369), deviating imaging (n=108), or missing clinical information (BMI only, n=29) were multiply imputed (5 imputation sets).

A second feature set was extracted using only images that are available in abbreviated breast MRI protocols. Processing step 2, lesion segmentation, and feature extraction were repeated using only the first post contrast images of the high-spatial-resolution dynamic image series. This feature set consisted of 36 image features and the same 3 clinical features.

Feature extraction was performed using MeVisLab (version 3.0, MeVis medical Solution AG, Bremen, Germany), Python (version 2.7, Python Software Foundation) with packages 'numpy' (v1.15.1) and 'scipy' (v1.1.0) and R (version 3.1.3, R Foundation for Statistical Computing, Vienna, Austria) with the packages 'psych' (v1.5.8) and 'Mice' (v2.25) was used for data imputation.

Training and validation

Outliers in feature values were defined as values deviating more than 3 standard deviations of the mean value. All feature values were normalized to values between 0 and 1. All BI-RADS 3, 4 and 5 lesions were used to train the model. The set was divided into ten folds, each fold contained 7 or 8 malignant and 40 or 41 benign lesions to maintain the prevalence of malignancy observed in the DENSE study. To train the prediction model, nine folds were used to fit a logistic regression model (the training set) the other fold was used as a validation set. BI-RADS 5 lesions were removed from the validation set, because in future application of the model, the gain of preventing false positive diagnosis is expected to be low in this group. Cross-validation was repeated 10 times, each fold was used as validation set once. Before model fitting, feature values labeled as outliers in the training set were censored by clipping the extreme values[26]. To prevent overfitting, model weights were reduced for each fit using Ridge regression[27]. By iterating over all folds, cross-validated probabilities were obtained for all lesions. The results of five imputation sets were combined using Rubin rules[28, 29]. The regularization parameter in the Ridge feature selection was determined using a second 10 fold cross-validation loop over the training data, using the deviance as performance

measure. The regularization parameter was selected 1 standard error above the parameter with lowest error. Hence, we chose the simplest model whose accuracy was comparable with the best model[30]. The posterior probabilities of the model to predict the presence of malignant disease in BI-RADS 3 and 4 lesions were summarized in an accuracy measure curve (AMC)[31] and receiver operator characteristic (ROC) curve. An AMC shows the percentage of correctly predicted malignant lesions and correctly predicted benign lesions for a range of values of the probability threshold (pt) using:

$$Sensitivity(pt) = \frac{TP(pt)}{TP(pt) + FN(pt)} \quad (4.1)$$

$$Specificity(pt) = \frac{TN(pt)}{TN(pt) + FP(pt)} \quad (4.2)$$

$$PPV(pt) = \frac{TP(pt)}{TP(pt) + FP(pt)} \quad (4.3)$$

$$NPV(pt) = \frac{TN(pt)}{TN(pt) + FN(pt)} \quad (4.4)$$

where TP is the number of true positives, TN the number of true negatives, FP the number of false positives and FN the number of false negatives for each probability threshold. In each decision curve, three operating points were selected, at a sensitivity level of 100%, 99% and 95% respectively. The models were compared using the McNemar chi-square test, a p-value less than 0.05 was considered significant.

Training and validation were performed using R (version 3.1.3, R Foundation for Statistical Computing, Vienna, Austria) with the packages 'psych' (v1.5.8), 'glmnet' (v2.0-5), and 'pROC' (v1.8)

4.3.4 Results

In the total screening population of 4783 women, 81 malignant lesions were found in 79 women and 444 benign lesions were found in 390 women. Fifteen women had both a malignant and a benign lesion. Four malignant lesions in three women were excluded because the images were not available. Fifteen benign lesions were excluded: Images of 9 lesions were not available for this study, one lesion was imaged using a deviant MRI protocol and five lesions could not be exam-

ined due to imaging artifacts affecting correct segmentation or feature extraction. Image artifacts were caused by movement of the woman during imaging. The deviant images or artifacts did not alter the ability of the radiologist to score the images; however, the images were unusable for the automated method presented. In summary, 429 benign and 77 malignant lesions were available for this study (Table 4.1, Table 4.2 and Table 4.3). Most lesions were scored BI-RADS 4: 293, while 194 lesions were scored BI-RADS 3 and 19 lesions were scored BI-RADS 5 (Table 4.2).

The accuracy measure curve and corresponding ROC curve (Figure 4.1) show feasibility to increase the specificity using CAD. An accuracy measure curve and ROC curve were obtained for the subgroup of all BI-RADS 3 and BI-RADS 4 lesions. The presented model outputs a probability of malignancy for each lesion. Cut-off thresholds in this probability define the sensitivity and specificity of the model. We chose three cut-off thresholds corresponding to sensitivity 100%, 99%, and 95%. The cut-off thresholds are shown in the accuracy measure curves (Figure 4.1). Corresponding specificity at each threshold is shown in Table 4.4.

The full multi-parametric model classified 176/425 (41.5%) of the false-positive BI-RADS 3 and BI-RADS 4 scored lesions as benign without missing a malignancy. Of the total group of lesions referred to additional MRI or biopsy, 425/487 (87.3%) were false-positive. With additional CAD used before referral, this fraction may be reduced to 311/487 (63.9%). For the abbreviated protocol model, the referrals would be 376 instead of 487 (77.2%). Examples of lesions that were false positive and correctly identified as such by computer aided diagnosis are shown in Figure 4.2 and Figure 4.3.

Table 4.1: Lesion types of BI-RADS 3, 4 and 5 lesions. Results were obtained from the pathology reports after biopsy.

Benign lesions	429
Adenomyoepithelioma	2
Adenosis	24
Apocrine metaplasia	14
Atypical ductal hyperplasia	5
Cholesterol crystal	1
Cylindrical cell metaplasia	1
Cyst	8
Fat necrosis	1
Fibroadenoma	40
Fibrosis	35
Hemangioma	2
LCIS*	4
Lipoma	3
Lobular hyperplasia	1
Lobular neoplasia	4
Lobulitis	3
Lymph node	5
Mastopathy	27
Normal breast tissue	23
Papilloma	18
Usual ductal hyperplasia	32
Unknown**	176
Malignant lesions	77
Ductal carcinoma in situ	13
Invasive ductal carcinoma	35
Invasive ductal lobular carcinoma	5
Invasive intracystic papillary carcinoma	2
Invasive lobular carcinoma	13
Invasive mucinous carcinoma	1
Invasive tubular carcinoma	8

*In the DENISE trial LCIS is considered a benign lesion[8], conform Dutch guidelines[32].

**No biopsy result was available for these lesions. No biopsy performed after BI-RADS 3 score (n=153), the lesion was not/no longer visible on additional imaging (n=16), biopsy result unknown(n=7)

Table 4.2: Overview of all lesions used for development of the CAD model stratified by BI-RADS score and Mass or Non Mass Enhancing lesions (NME).

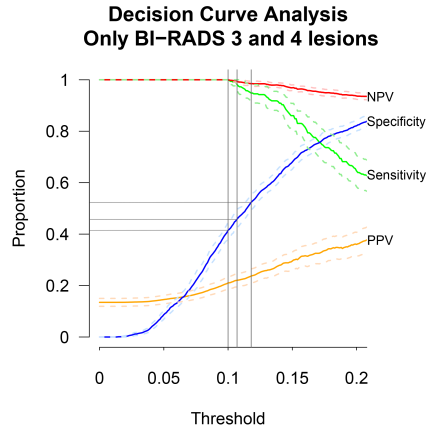
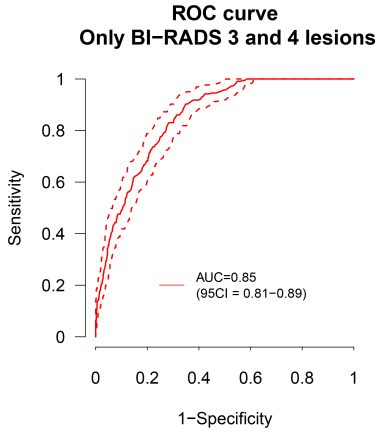
BI-RADS score		Number of benign lesions	Number of malignant lesions
3	Mass	105	3
	NME	84	2
4	Mass	168	47
	NME	68	10
5*	Mass	4	14
	NME	0	1
total		429	77

*only used for model training, not for model testing.

Table 4.3: Comparison between malignant and benign lesions used for this study; median feature values and interquartile range are shown. For statistical comparison the Kruskal Wallis test was used, $p < 0.05$ was considered significant.

	Benign	Malignant	
Number of lesions	429	77	
Lesion volume (cm³)	0.18 (0.10-0.36)	0.33 (0.16-0.77)	$p < 0.001$
BMI	22.25 (20.75-24.01)	22.86 (21.48-24.95)	$p = 0.013$
Age (years)	53.05 (50.90-56.90)	54.80 (51.30-61.70)	$p = 0.008$

Full protocol



Abbreviated protocol

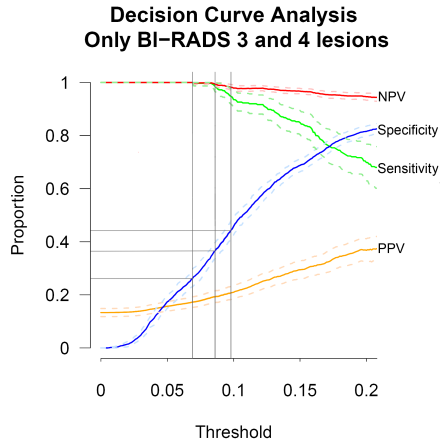
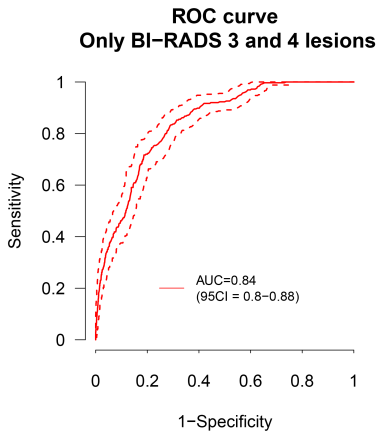


Figure 4.1: Accuracy measure curve of the CAD model and corresponding ROC curve using all MRI imaging series (top row) and abbreviated imaging series (bottom row). The blue curve denotes specificity, the yellow curve the positive predictive value (PPV), the green curve the sensitivity and the red curve the negative predictive value (NPV). One standard deviation corrected for multiple imputation using Rubin rules is shown using dashed curves. The vertical gray lines indicate the cut-off threshold probabilities (pt) corresponding to sensitivity 100%, 99.0% and 95.0% from left to right.

Table 4.4: Overview of correctly classified benign BI-RADS 3 and 4 lesions and corresponding levels of correctly classified malignant lesions for both models. Results denote mean \pm 1 standard deviation. Models were compared using the McNemar chi-square statistic.

Correctly classified malignant lesions	Correctly classified benign BI-RADS 3 and 4 lesions		
	Full protocol	Abbreviated protocol	
100.0%	41.5% \pm 3.2%	26.2% \pm 3.2%	p<0.01
99.0%	45.8% \pm 3.5%	36.6% \pm 3.0%	p<0.01
95.0%	52.4% \pm 3.1%	44.3% \pm 3.6%	p<0.01

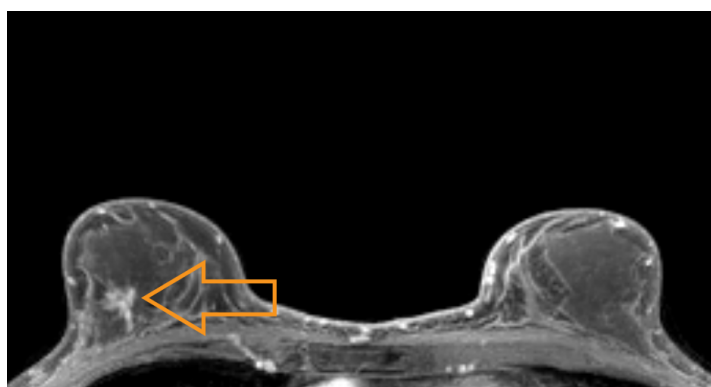


Figure 4.2: Maximum intensity projection of a 14 mm false-positive lesion in a 59 year old woman who was referred to biopsy. The BI-RADS 4 classified lesion (right breast) was a benign fibrotic lesion (arrow). The computer-aided diagnosis correctly classified it as benign with probability of malignancy of 2.5%.

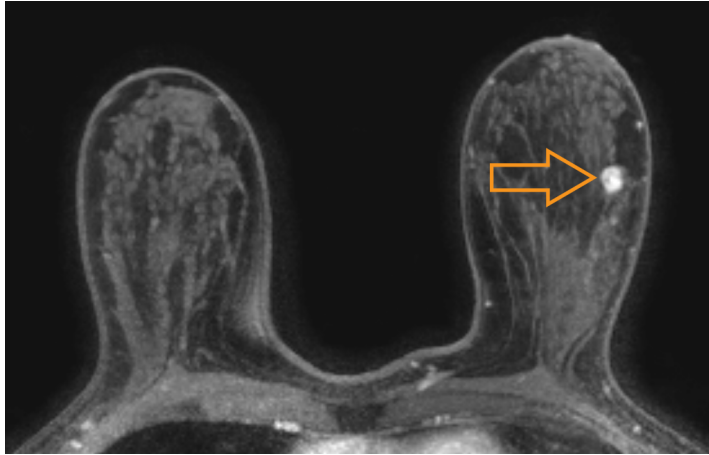


Figure 4.3: Maximum intensity projection of a 10 mm false-positive lesion in a 52 year old woman who was referred to biopsy. The BI-RADS 4 classified lesion (left breast) was a benign fibroadenoma (arrow). The computer-aided diagnosis correctly classified it as benign with probability of malignancy of 2.1%.

4.4 Discussion

From a screening population of 4783 women, MR images of 77 malignant breast lesions and 429 benign lesions were used to create a multi-parametric CAD model based on Ridge regression, to identify benign disease with high certainty. The model may have potential to reduce follow up on benign BI-RADS 3 and BI-RADS 4 lesions: 41.5% reduction for the full multi-parametric protocol and 26.2% for the abbreviated protocol model, without missing a malignant lesion.

Although the performance of the full-protocol model and the abbreviated-protocol model is comparable in terms of AUC (0.85 vs 0.84), the number of detected benign lesions without missing malignant lesions was observed to be higher in the full-protocol model ($p < 0.01$). These results suggest that the high-resolution post contrast images contain information to increase the specificity to identify benign disease at high sensitivity. We observed comparable performance between mass and non-mass lesions, indicating that the features accurately describe both types of lesions (Supplemental Material 2: Performance in mass and non-mass enhancing lesions).

To our knowledge, this study is the first to apply a multi-parametric CAD model in unselected homogeneous data obtained from a multicenter screening trial in women with extremely dense breasts. The performance of presented models,

(AUC of 0.85 ± 0.04 and 0.84 ± 0.04) based on DCE, fast-DCE, ADC, T2 and clinical data, is on par to that of other published methods in other study populations. Dalmış et al. used deep learning on fast-DCE, T2 and DWI and obtained an AUC of 0.852[9]. Other authors designed multi-parametric models using fewer image sequences for feature extraction, e.g. DCE and T2-weighted images[11, 33, 34], resulting in AUCs of 0.88 ± 0.01 [11], 0.83 ± 0.03 [33] or 0.85 ± 0.03 [34]. Using only DCE yielded comparable results (AUC of 0.85)[35]. However, the above results might not be directly comparable to our results because they were based on single institution data. In addition, the current study omitted BI-RADS 2 and 5 lesions. The rationale for this omission is that the problem of false positives does not occur in BI-RADS 2 and BI-RADS 5. BI-RADS 2 lesions are not referred and by definition, BI-RADS 2 lesions are nearly always benign. BI-RADS 5 lesions are nearly always malignant. By omitting these categories, the CAD is tested on the most difficult and clinically relevant cases.

In this study, the risk of overfitting was reduced using Ridge regression[36]. Features with the largest regression weights were signal enhancing ratio, top washout, volume uptake and volume washout from the DCE image series, the maximum slope and general slope from the fast-DCE, and the 75th percentile of ADC values in the lesion. T2 features did not have high weighting in the model.

We did not use deep learning because the number of malignant lesions was relatively small for such an approach. Deep learning can outperform linear regression methods when the number of training data is large enough to avoid overfitting[11]. Currently, however, the literature does not indicate a clear benefit of deep learning over radiomics for this problem, other than that deep-learning models are less time-consuming to construct. The largest study on deep learning to discriminate between benign and malignant disease on MRI uses 1294 cases[11], and yields comparable performance (AUC of 0.88). A potential risk of deep learning is, however, that the millions of parameters that describe the data may cause unnoticed bias in the detection of malignant disease in populations for which the model was not explicitly trained. Although all MRI data were acquired according to the same protocol, variation was introduced between institutions because MRI scanners from different vendors were used. Moreover, some MRI settings varied across hospitals, e.g., the use of fat suppression, flip angle and repetition time. We used a data harmonization step between MRI scanners to counter the effect of some of these variations.

This study also has some limitations. We were not able to validate the method

in an unseen dataset, but used cross-validation. In future research, the CAD model should be validated in an independent population of women with extremely dense breasts. Another potential limitation is that we used lesions found only during the first round of the screening trial. We have not yet investigated whether the machine-extracted phenotype of lesions detected in subsequent, or incident, screening rounds is representative of that detected in the first round. For instance, lesions may be smaller on average in subsequent screening rounds, and perhaps also more aggressive. The tumors in the first, or prevalent round, may comprise of relatively slow-growing, less aggressive tumors that have been present for a long time. We describe computerized analysis of MR images with BI-RADS score of the radiologist as input. CAD is, however, typically implemented as an aid to radiologists, using the computer as second opinion. This interaction has not yet been investigated.

In conclusion, we developed a CAD method based on ridge regression to identify benign lesions with high certainty in multi-parametric breast MR screening of extremely dense breasts, thus pursuing to reduce the number of recalls on benign lesions. Using internal validation, the method showed potential to reduce referral of benign BI-RADS 3 and BI-RADS 4 lesions without loss of sensitivity.

References

- [1] P. A. Carney, D. L. Miglioretti, B. C. Yankaskas, and et al., "Individual and combined effects of age, breast density, and hormone replacement therapy use on the accuracy of screening mammography," *Annals of Internal Medicine*, vol. 138, no. 3, pp. 168–175, 2003.
- [2] W. A. Berg et al., "Detection of breast cancer with addition of annual screening ultrasound or a single screening mri to mammography in women with elevated breast cancer risk," *JAMA*, vol. 307, no. 13, pp. 1394–404, 2012.
- [3] C. K. Kuhl et al., "Abbreviated breast magnetic resonance imaging (mri): first postcontrast subtracted images and maximum-intensity projection—a novel approach to breast cancer screening with mri," *J Clin Oncol*, vol. 32, p. 2304, 2014.
- [4] C. K. Kuhl et al., "Supplemental breast mr imaging screening of women with average risk of breast cancer," *Radiology*, vol. 283, no. 2, pp. 361–370, 2017.
- [5] S. Saadatmand et al., "Mri versus mammography for breast cancer screening in women with familial risk (famrisc): a multicentre, randomised, controlled trial," *The Lancet Oncology*, 2019.
- [6] G. L. Menezes et al., "Magnetic resonance imaging in breast cancer: A literature review and future perspectives," *World journal of clinical oncology*, vol. 5, no. 2, pp. 61–70, 2014.
- [7] M. J. Emaus et al., "Mr imaging as an additional screening modality for the detection of breast cancer in women aged 50-75 years with extremely dense breasts: The dense trial study design," *Radiology*, vol. 277, no. 2, pp. 527–37, 2015.
- [8] M. F. Bakker et al., "Supplemental mri screening for women with extremely dense breast tissue," *New England Journal of Medicine*, vol. 381, no. 22, pp. 2091–2102, 2019.
- [9] M. U. Dalmiş et al., "Artificial intelligence–based classification of breast lesions imaged with a multiparametric breast mri protocol with ultrafast dce-mri, t2, and dwi," *Investigative Radiology*, vol. Publish Ahead of Print, 2019.
- [10] M. Zhang et al., "Multiparametric mri model with dynamic contrast-enhanced and diffusion-weighted imaging enables breast cancer diagnosis with high accuracy," *Journal of Magnetic Resonance Imaging*, vol. 0, no. 0, 2019.
- [11] D. Truhn et al., "Radiomic versus convolutional neural networks analysis for classification of contrast-enhancing lesions at multiparametric breast mri," *Radiology*, vol. 290, no. 2, pp. 290–297, 2019.
- [12] K. G. A. Gilhuijs et al., "Breast mr imaging in women at increased lifetime risk of breast cancer: Clinical system for computerized assessment of breast lesions—initial results," *Radiology*, vol. 225, no. 3, pp. 907–916, 2002.
- [13] M. U. Dalmiş et al., "A computer-aided diagnosis system for breast dce-mri at high spatiotemporal resolution," *Medical Physics*, vol. 43, no. 1, pp. 84–94, 2016.

- [14] A. Tahmassebi et al., "Impact of machine learning with multiparametric magnetic resonance imaging of the breast for early prediction of response to neoadjuvant chemotherapy and survival outcomes in breast cancer patients," *Investigative Radiology*, vol. 54, no. 2, pp. 110–117, 2019.
- [15] Y. Kuroki et al., "Diffusion-weighted imaging of breast cancer with the sensitivity encoding technique: Analysis of the apparent diffusion coefficient value," *Magnetic Resonance in Medical Sciences*, vol. 3, no. 2, pp. 79–85, 2004.
- [16] L. Ballesio et al., "Breast mri: Are t2 ir sequences useful in the evaluation of breast lesions?," *European Journal of Radiology*, vol. 71, no. 1, pp. 96–101, 2009.
- [17] C. M. Chhor and C. L. Mercado, "Abbreviated mri protocols: Wave of the future for breast cancer screening," *American Journal of Roentgenology*, vol. 208, no. 2, pp. 284–289, 2016.
- [18] A. Gubern-Mérida et al., "Automated localization of breast cancer in dce-mri," *Medical Image Analysis*, vol. 20, no. 1, pp. 265–274, 2015.
- [19] T. Alderliesten et al., "Validation of semiautomatic measurement of the extent of breast tumors using contrast-enhanced magnetic resonance imaging," *Investigative Radiology*, vol. 42, no. 1, pp. 42–49, 2007.
- [20] E. M. Haacke et al., "New algorithm for quantifying vascular changes in dynamic contrast-enhanced mri independent of absolute t1 values," *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, vol. 58, no. 3, pp. 463–472, 2007.
- [21] M. J. van Rijssel et al., "Correcting time-intensity curves in dynamic contrast-enhanced breast mri for inhomogeneous excitation fields at 7t.," *Magnetic Resonance in Medicine*, vol. In Press, 2019.
- [22] E. Verburg et al., "Knowledge-based and deep learning-based automated chest wall segmentation in magnetic resonance images of extremely dense breasts," *Medical physics*, 2019.
- [23] G. S. Chilla, C. H. Tan, C. Xu, and C. L. Poh, "Diffusion weighted magnetic resonance imaging and its recent trend-a survey," *Quantitative imaging in medicine and surgery*, vol. 5, no. 3, pp. 407–422, 2015.
- [24] S. Klein et al., "A toolbox for intensity-based medical image registration," *IEEE Transactions on Medical Imaging*, vol. 29, no. 1, pp. 196–205, 2010.
- [25] K. G. A. Gilhuijs, M. L. Giger, and U. Bick, "Computerized analysis of breast lesions in three dimensions using dynamic magnetic-resonance imaging," *Medical Physics*, vol. 25, no. 9, pp. 1647–1654, 1998.
- [26] C. Hastings, F. Mosteller, J. W. Tukey, and C. P. Winsor, "Low moments for small samples: A comparative study of order statistics," *Ann. Math. Statist.*, vol. 18, no. 3, pp. 413–426, 1947.

- [27] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970.
- [28] A. Marshall, D. G. Altman, R. L. Holder, and P. Royston, "Combining estimates of interest in prognostic modelling studies after multiple imputation: current practice and guidelines," *BMC medical research methodology*, vol. 9, pp. 57–57, 2009.
- [29] D. Rubin, *Multiple Imputation for Nonresponse in Surveys*. Wiley, 2004.
- [30] L. Breiman, *Classification and regression trees*. Wadsworth International Group, 1984.
- [31] S. Vos et al., "Comprehensive proteomic profiling–derived immunohistochemistry-based prediction models for brca1 and brca2 germline mutation-related breast carcinomas," *The American journal of surgical pathology*, vol. 42, no. 9, pp. 1262–1272, 2018.
- [32] Oncoline, "Borstkanker - algemeen," 2017.
- [33] C. Gallego-Ortiz and A. L. Martel, "Using quantitative features extracted from t2-weighted mri to improve breast mri computer-aided diagnosis (cad)," *PLOS ONE*, vol. 12, no. 11, p. e0187501, 2017.
- [34] N. Bhooshan et al., "Combined use of t2-weighted mri and t1-weighted dynamic contrast-enhanced mri in the automated analysis of breast lesions," *Magnetic resonance in medicine*, vol. 66, no. 2, pp. 555–564, 2011.
- [35] E. E. Deurloo et al., "Clinically and mammographically occult breast lesions on mr images: Potential effect of computerized assessment on clinical reading," *Radiology*, vol. 234, no. 3, pp. 693–701, 2005.
- [36] M. van Smeden et al., "Sample size for binary logistic prediction models: Beyond events per variable criteria," *Statistical Methods in Medical Research*, vol. 0, no. 0, p. 0962280218784726, 2018.

4.5 Supplemental Material 1: Description of image features

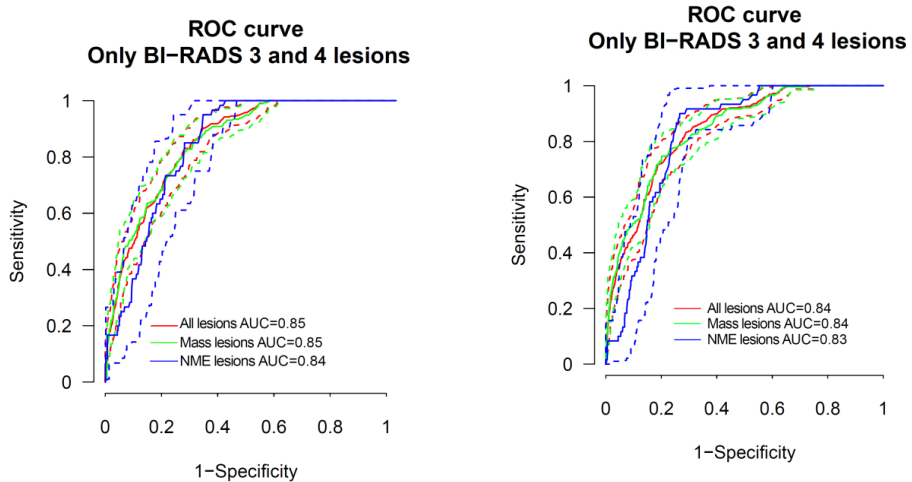
Table 4.5: Overview of imaging features with a short description for each feature. Features are grouped by MRI sequence.

DCE[25, 12]			DCE (continued)		
1	Circularity	Measure for how similar the tumor shape is to a sphere	20	Mean smoothness	Maximum mean radial gradient intensity
2	Irregularity	Measure for the roughness of the tumor surface	21	Standard deviation Radial gradient histogram analysis frame 2*	Standard deviation radial gradient histogram analysis at time point with maximum value
3	Volume	Volume of the tumor	22	Radial gradient histogram analysis frame 2*	Radial gradient histogram analysis at time point with maximum value
4	Largest diameter	Largest distance between voxels pairs in the tumor segmentation	Fast-DCE[13]		
5	Uptake	Average of (I1-I0)/I0 over the tumor voxels; I0 and I1 are the precontrast and the first post-contrast signal intensities	1	Maximum slope	Maximum slope of uptake contrast agent in lesion volume
6	Washout*	Average of (I4-I1)/I1 over the tumor voxels; I1 and I4 are the first and last postcontrast signal intensities	2	Time of maximum slope	Time between maximum slope of contrast uptake in descending aorta and lesion volume
7	Signal enhancing ratio (SER)*	Average of (I1-I0)/(I4-I0) over the tumor voxels; I0, I1 and I4 are the, precontrast and first and last postcontrast signal intensities	3	Time to enhancement	Time between maximum contrast uptake in descending aorta and start contrast uptake in lesion volume
8	Top uptake	Average uptake of the top 10 percent enhancing tumor voxels	4	Washout	Intensity gradient at last time point of Fast-DCE
9	Top washout*	Average washout of the top 10 percent enhancing tumor voxels	5	General slope	Maximal slope of contrast uptake in lesion between time point aorta and any other time point during contrast uptake.
10	Volume uptake	Volume of tumor in washin image	6	Maximum enhancement	Maximal normalized intensity in lesion volume
11	Largest diameter uptake	Largest diameter of tumor in washin image	T2		
12	Volume washout*	Volume of tumor in washout image	1	Minimum intensity	All T2 intensities are normalized to the intensity of the pectoral muscle(15) Minimum intensity in lesion volume
13	Largest diameter washout*	Largest diameter of tumor in washout image	2-8	5th, 10th, 25th, 50th, 75th, 90th and 95th percentile	Percentile of the intensities present in the lesion volume
14	Mean sharpness / margin gradient	The sharpness of the uptake of contrast at the tumor margin	9	Maximum intensity	Maximal intensity in lesion volume
15	Variance of sharpness / variance of margin gradient	The variance in sharpness of uptake of contrast at the tumor margin	ADC		
16	Variation sharpness	Variance of sharpness at time point with maximum mean sharpness	1	Minimum intensity	Minimum intensity in lesion volume
17	Mean sharpness frame 2*	Mean sharpness at first post-contrast	2-8	5th, 10th, 25th, 50th, 75th, 90th and 95th percentile	Percentile of the intensities present in the lesion volume
18	Variance sharpness frame 2*	Variation sharpness at first post-contrast	9	Maximum intensity	Maximal intensity in lesion volume
19	Variation smoothness	Maximum standard deviation radial gradient histogram (RGH) values, see			

*only used for model training, not for model testing.

4.6 Supplemental Material 2: Performance in mass and non-mass enhancing lesions

The performance of both the full protocol model and the abbreviated protocol model showed no differences between mass and non-mass enhancing (NME) lesions.



(a) ROC curve of the full protocol model. In red the performance on all lesions, in green on the subset of mass enhancing lesions and in blue the performance of the model on NME lesions

(b) ROC curve of the abbreviated protocol model. In red the performance on all lesions, in green on the subset of mass enhancing lesions and in blue the performance of the model on NME lesions

For more detail we also show the DCA of the different lesion subsets. As shown in the Table 4.6 and Table 4.7, the performance measure of correctly classified benign BI-RADS 3 and 4 lesions is comparable to the overall performance of the model.

Table 4.6: DCA curve results of different subgroups of lesions, all lesions (top row), mass lesions (middle row) and NME lesions (bottom row) for both proposed models

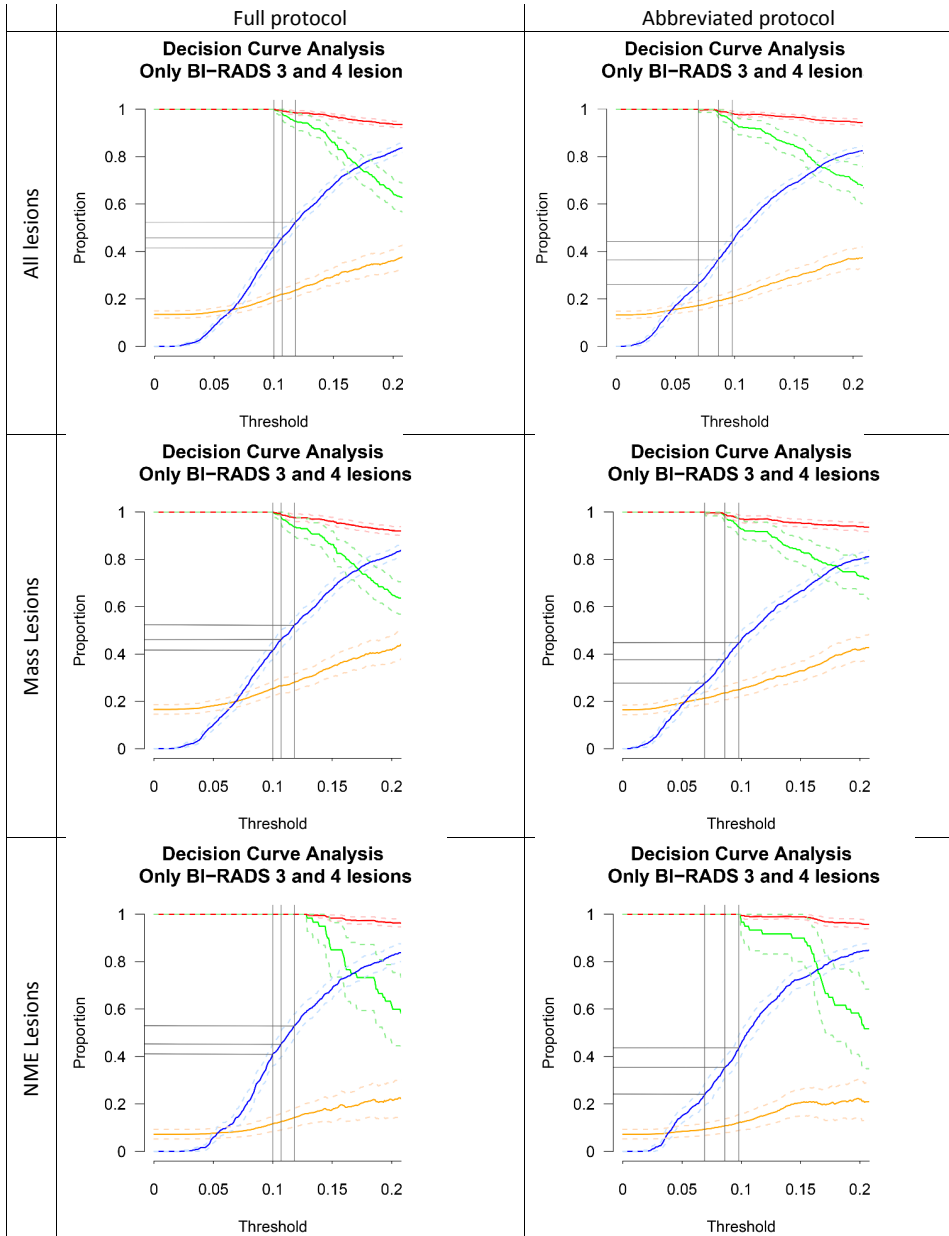


Table 4.7: Overview of correctly classified benign BI-RADS 3 and 4 lesions for the subset of lesions and corresponding levels of correctly classified malignant lesions for both models. Results denote mean \pm 1 standard deviation. Models were compared using the McNemar chi-square statistic.

Correctly classified lesions	Correctly classified benign BI-RADS 3 and 4 lesions		
	Full protocol	Abbreviated protocol	
	All lesions		
100.0%	41.5% \pm 3.2%	26.2% \pm 3.2%	p<0.01
99.0%	45.8% \pm 3.5%	36.6% \pm 3.0%	p<0.01
95.0%	52.4% \pm 3.1%	44.3% \pm 3.6%	p<0.01
	Mass lesions		
	Full protocol	Abbreviated protocol	
100.0%	41.6% \pm 3.7%	27.6% \pm 3.6%	p<0.01
99.0%	46.1% \pm 3.8%	37.3% \pm 3.4%	p<0.01
95.0%	52.1% \pm 3.4%	44.7% \pm 3.6%	p<0.01
	NME lesions		
	Full protocol	Abbreviated protocol	
100.0%	41.2% \pm 3.2%	23.8% \pm 4.8%	p<0.01
99.0%	45.4% \pm 5.5%	35.3% \pm 4.6%	p<0.01
95.0%	52.8% \pm 5.1%	43.7% \pm 5.5%	p<0.01

Chapter 5

Validation of combined Deep-learning Triaging and Computer-aided Diagnosis in 2,901 Breast MRI Examinations from the Second Screening Round of the DENSE Trial

Based on: **Erik Verburg**, Carla H. van Gils, Bas H.M. van der Velden, Marije F. Bakker, Ruud M. Pijnappel, Wouter B. Veldhuis, Kenneth G. A. Gilhuijs; Validation of combined Deep-learning Triaging and Computer-aided Diagnosis in 2,901 Breast MRI Examinations from the Second Screening Round of the DENSE Trial, *Accepted for publication in Investigative Radiology, 2023*

5.1 Abstract

Background

Computer-Aided Triaging (CAT) and Computer-Aided Diagnosis (CAD) of screening breast MRI have shown potential to reduce the workload of radiologists in the context of dismissing normal breast scans and dismissing benign disease in women with extremely dense breasts.

Purpose

To validate the potential of integrating CAT and CAD to reduce workload and workup on benign lesions in the second screening round of the DENSE trial, without missing cancer.

Methods

We included 2901 breast MRI scans, obtained from eight hospitals in the Netherlands. CAT and CAD were previously developed on data from the first screening round. CAT dismissed examinations without lesions. MRI examinations triaged to radiological reading were counted and subsequently processed by CAD. The number of benign lesions correctly classified by CAD was recorded. The false-positive fraction of the CAD was compared with that of unassisted radiological reading in the second screening round. Receiver Operating Characteristics (ROC) analysis was performed and the generalizability of CAT and CAD were assessed by comparing results from first and second screening rounds.

Results

CAT dismissed 950/2901 (32.7%) examinations with 49 lesions in total, none were malignant. Subsequent CAD classified 132/285 (46.3%) lesions as benign without misclassifying any malignant lesion. Together, CAT and CAD yielded significantly less false-positive lesions, 53/109 (48.6%) vs 89/109 (78.9%) ($p=0.001$) than radiological reading alone. CAT had smaller area under the ROC curve (AUC) in the second screening round compared to the first, 0.83 versus 0.76 ($p=0.001$). CAD was not associated with significant differences in AUC (0.857 versus 0.753, $p=0.08$). At the operating thresholds the performances of CAT (39.7% versus

41.0%, $p=0.70$) and CAD (41.0% versus 38.2%, $p=0.62$) were successfully reproduced in the second round.

Conclusion

The combined application of CAT and CAD showed potential to reduce workload of radiologists and to reduce number of biopsies on benign lesions.

5.2 Introduction

Contrast-enhanced MRI may be used in combination with X-ray mammography to screen asymptomatic women for breast cancer. Supplemental MRI screening in women with extremely dense breasts improved the detection of cancer[1]. Similar observations were reported for women at increased life-time risk. Nonetheless, breast MRI screening has lower specificity compared to mammography[1, 3, 4] and it invokes additional workload.

To reduce the workload of breast MR radiologists, researchers have focused on automated lesion detection[5, 6]. One focused on identifying normal scans using Computer-Aided-Triaging (CAT)[7]. Computer-Aided Diagnosis (CAD) of dynamic contrast-enhanced MRI[8, 9] and multi-parametric MRI[1, 10] were found to further increase specificity[11–15].

A recently reported CAT - developed on data from 4 783 MRI examinations from the first screening round of the DENSE trial - dismissed approximately 40% of normal breast examinations without dismissing malignant disease[7]. In addition to CAT, CAD was developed on the same data to distinguish between 444 benign and 81 malignant lesions. It is yet unknown whether CAD is complementary to CAT to increase the positive-predictive value of MRI screening in women with extremely dense breasts while maintaining high negative-predictive value, and minimizing the number of normal scans to be read by radiologists.

The aim of this study is to validate the potential of combining CAT with CAD in the second screening round of DENSE to minimize work load as well as minimizing the number of biopsies on benign lesions without dismissing malignant breast disease.

5.3 Materials and Methods

We validate the potential impact of combined CAT and CAD in the second screening round of the DENSE trial, and compare it to radiological reading without computer assistance. Impact is expressed in terms of reduction in 1) MRI scans with normal anatomy read by radiologists; 2) false-positive referrals to further diagnostic work-up with additional MRI or biopsy. Both CAT and CAD were previously trained on MRI scans from the first screening round only.

First, we briefly describe the design of the DENSE trial, followed by description of the study participants, MRI acquisition parameters, unassisted radiological read-

ing (i.e., the reference standard), CAT and CAD, followed by the combination of the methods.

5.3.1 DENSE trial

The DENSE trial (ClinicalTrials.gov: NCT01315015) investigates whether additional MRI-screening of asymptomatic women with extremely dense breasts (i.e., ACR BI-RADS category 4 measured with Volpara software) reduces the number of interval cancers [16]. Participating women had extremely dense breasts without lesions suspected of malignancy on mammography. The first results of the DENSE-trial confirmed the hypothesis of detection of additional breast cancers and the reduction of interval cancers. In the first round of screening, the cancer-detection yield with MRI after negative mammography was 79 in 4783 women, or 16.5/1000 screens[1].

The current validation study focused primarily on the screening data from the second round. No prior AI studies have been performed on these data before. The screened data acquired in the first round were included in two prior AI studies, one on AI -triaging[7], the other on computer-aided diagnosis[15].

5.3.2 participants

Participants (between 50 and 75 years of age) were included from the national population-based mammography screening program. From the 4783 participants in the first MRI screening round of DENSE, 3436 women participated in the second MRI round between 6 September 2014 and 17 April 2019 [17]. To be eligible for the second MRI round they had been participating in the national program, again with a normal mammography result (i.e., no referral). Written informed consent was obtained from all women before screening. The trial was approved by the Dutch Minister of Health, Welfare and Sport (2011/19 WBO, The Hague, the Netherlands). According to the Dutch law on population studies, the study was waived from ethical review by the local institutional review board.

5.3.3 MRI Acquisition

MRI examinations were performed in eight hospitals in the Netherlands using the same MRI protocol in each screening round. The protocol has been described in detail elsewhere[16]. In short, T1-weighted images were acquired without fat

suppression, followed by dynamic T1-weighted imaging, consisting of one pre contrast series at high spatial resolution and 15 to 20 fast acquisitions after contrast administration. Four to five post-contrast series at high spatial resolution followed. Fat suppression was optional. In addition, diffusion-weighted series were acquired using two or three b-values. T2-weighted acquisition was optional. Contrast agent was injected at rate of 1 mL/sec to a total dose of 0.1-mmol of gadobutrol (Gadovist; Bayer AG, Leverkusen, Germany) per kilogram of body weight. Images were acquired using a 3-Tesla MRI unit; five hospitals used Philips MR devices (Eindhoven, the Netherlands), the other three hospitals used Siemens devices (Erlangen, Germany).

5.4 Methods

5.4.1 Unassisted radiological reading

In the DENSE trial, breast MR examinations were read by trained breast MR radiologists (with experience from 5 to 23 years[1]). In short, MRI examinations were single read, and scored according to the BI-RADS MRI lexicon [18]. Only BI-RADS 3 lesions were double read (consensus reading), in these cases MRI was repeated after 6 months. Women with BI-RADS 4 or BI-RADS 5 lesions were always recommended to undergo biopsy.

5.4.2 CAT

The method previously developed[7] to dismiss the largest number of normal breast MRI examinations without dismissing malignant disease was applied, without modifications, to the second screening round. In short, the probability of lesion presence was estimated using deep learning. This was done for each breast separately. The probability was established in three MIP images of contrast-agent uptake in orthogonal directions (transversal, sagittal and coronal), and the three results were averaged. The probability per examination was equal to the highest probability in the left or right breast.

During model development on first screening round data, eight-fold internal-external validation was used, i.e., in each fold, data of one hospital were hold out as test data and the data of the remaining hospitals were used to train the convolutional neural network (CNN). Hence, eight CNNs were developed (one for each fold). When the probability of lesion presence was less than an operating threshold

(established in the first screening round data), the breast examination was considered normal[7].

5.4.3 CAD

Previously, a method was developed to distinguish between benign and malignant breast lesions on multi-parametric MRI[15]. In short, lesion segmentation was followed by feature extraction and classification into benign or malignant groups. Lesion segmentation used constrained volume growing from a manually placed seed point[19] at or near the lesion by a technical physician (E.V.) under supervision of a breast MR radiologist (W.V.). The features were extracted from the segmentation results and the MR images. In addition, clinical features were used (i.e., age, BMI, and BIRADS)[15]. Training and testing was initially done on the first-round screening data only using Ridge-regression modelling with 10-fold cross validation to estimate probability of malignancy. In the current study, we retrained the Ridge-regression model on the first-round data and applied the model to the second-round screening data. An operating threshold in the probability was chosen in first screening round data at which all malignant lesions were correctly identified.

5.4.4 Combination of computer assisted triaging and computer-aided diagnosis

The current validation study applied CAT and CAD to the second round (Figure 5.1), using the operating points established in the first round[15, 7]. Scans considered to be normal by CAT were recorded and dismissed from further analysis. Scans considered to contain lesions were matched against the lesions detected by radiologists in the second screening round of the trial. These lesions were then offered to the CAD. Lesions considered to be benign and those considered to be malignant were recorded.

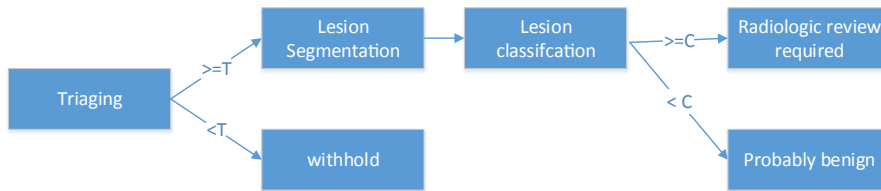


Figure 5.1: Combination of computer-aided triaging and computer-aided diagnosis applied to the second screening round of the DENISE trial. Breasts with probability of lesions lower than operating threshold T were dismissed by CAT for processing by CAD. If the probability of malignant disease was larger or equal to operating threshold C , the lesion was classified as malignant.

5.4.5 Avoiding bias

Because all participants in the second screening round were also screened in the first round, bias may occur when the round-1 model is validated on the round-2 data. Hence, the round-1 model was retrained on the round-1 data to avoid such bias, following the steps outlined below. To train the CAT model on the data from the first screening round, internal-external validation was used, meaning that the model was trained on data from seven hospitals and tested on data from the eighth hospital, alternating such that each hospital was used once as an external test set. Internal-external validation thus yielded eight models, and each model was constructed without overlap of women in training and test set. The overall performance of CAT in the first screening round was then estimated by averaging the performance of the eight models. To validate the CAT on the data from the second screening round, we used the same eight internal-external validation folds from the first round. That is, each of the eight CAT models from the first round was applied - without additional training - to the corresponding hospital in the second round. Hence, no woman in the first or second round was ever assessed by a model that included their training data. Again, the overall performance of the CAT in the second round was then assessed by averaging the performance of the eight models. To validate the CAD, we also took measures to avoid overlap in women in the training and test set: for each lesion detected in the second screening round, the first-round training data of that patient was removed from the CAD model before it was applied to the second-round screening data. Although BIRADS scores of radiologists are used by the CAD[15], these scores were ignored in this combined CAT/CAD assessment to mimic prospective autonomous application where radiologists have not yet assigned BIRADS scores. For this purpose, the

CAD model was established for each hospital separately using the first-round screening data without the BIRADS scores.

5.4.6 Statistics

The performance of combined CAT and CAD was compared with that of unassisted radiological reading during the DENSE trial. This was done as follows: CAT either detects no lesions in an examination or lesions in one or both breasts. These occurrences were counted separately, but the results are presented at examination level, i.e., whether one or more lesions are present in both breasts in the same examination according to the CAT. In addition, the number of correctly classified benign lesions by the CAD were counted. The false-positive rate of the CAD was compared with the false-positive rate of unassisted radiological reading (i.e., the rate of recalled suspicious lesions that turned out to be benign) using McNemar tests. A p-value of less than 0.05 was considered statistically significant.

The reproducibility of CAT and CAD separately were established by comparing the results from the first and second screening round. For this purpose, differences in area under the receiver operating characteristic (ROC) curve (AUC) were tested using the paired student t test (8 CAT models) or deLong test (1 CAD model). The percentage of examinations dismissed and the percentage of examinations with lesions that would be offered to radiologists by CAT were recorded and compared using paired student t test.

The CAD developed on the first round, was applied to BI-RADS 3, 4 and 5 lesions in round 2. In addition to AUC, Positive Predictive Value (PPV) and percentage of correctly classified benign lesions were compared between rounds using McNemar tests. It was verified that the negative predictive value (NPV) of CAT and CAD for malignant disease is 100%, as established in the first screening round.

5.5 Results

5.5.1 DENSE trial and unassisted radiological reading

In total, 2 901 (84.4%) MRI examinations of 3 436 women in the second screening round were included. Five-hundred-thirty-five women were excluded because their data could not be retrieved in full from participating hospitals. Unassisted by CAT or CAD, radiologists reported 334 lesions in 303 (of 2901) women. Three

women had three lesions and 25 had two lesions. Twenty lesions were malignant, 314 were benign (Table 5.1). The lesions were scored BI-RADS 2 (n=225), BI-RADS 3 (n=21), BI-RADS 4 (n=82) and BI-RADS 5 (n=6).

Table 5.1: Type of lesions in second screening round. Results were obtained from biopsy.

Benign lesions	314
Adenosis	2
Apocrine metaplasia	3
Atypical ductal hyperplasia	2
Cylindrical cell metaplasia	1
Cyst	3
Epithelia proliferation	1
Fibroadenoma	5
Fibrosis	8
Hemangioma	1
LCIS*	1
Lymph node	2
Mastopathy	4
Normal breast tissue	3
Papilloma	3
Periductitis	1
Sclerosis	5
Usual ductal hyperplasia	7
BI-RADS 2 (no biopsy)	225
BI-RADS 3 (no biopsy)	21
Unknown	16
Malignant lesions	20
Ductal carcinoma in situ	6
Invasive carcinoma (not otherwise specified)	8
Mixed invasive ductal and lobular carcinoma	2
Invasive lobular carcinoma	3
Invasive tubular carcinoma	1

*In the DENISE trial LCIS is considered a benign lesion[1], conform Dutch and international guidelines[20].

5.5.2 CAT

The performance of CAT in the second screening round is shown in Table 5.2. CAT showed a smaller area under the ROC curve in the second screening round than in the first screening round (0.76 versus 0.83, $p=0.001$) (Figure 2). We found no evidence of differences in performance at the operating threshold ($p=0.70$). In the second round 41.0% (95% CI: 30.4 - 51.6) of the examinations without any lesions would be dismissed compared to the 39.7% (95% CI: 30.0 - 49.4) in the first screening round[7]. The percentage of examinations with lesions that would continue to radiological review was also not different ($p=0.07$) in the second screening round (85.6% [95% CI: 79.2 - 92.0] versus 90.7% [95% CI: 86.7 - 94.7]) in the first screening round. No examinations with malignant disease were dismissed, i.e., NPV=100%.

Table 5.2: Results of triaging in first and second round data.

	First-round data	Second- round data	p-value
AUC	0.83 (0.80-0.85)	0.76 (0.72-0.81)	$p=0.001$
Percentage of dismissed examinations without lesion	39.7 (30.0-49.4)	41.0 (30.4-51.6)	$p=0.70$
Percentage of examinations with lesions triaged to radiological review	90.7 (86.7-94.7)	85.6 (79.2-92.0)	$p=0.07$

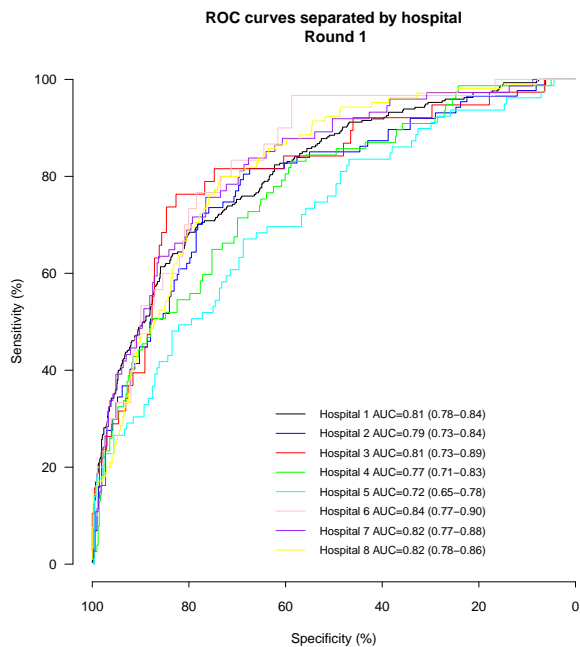


Figure 5.2: ROC curves of Computer-Aided Triaging for the task of distinguishing between examinations with lesions (benign and malignant) and examinations without lesions, applied to first (left) and second (right) screening-round data. The 95% confidence intervals are shown in the legend.

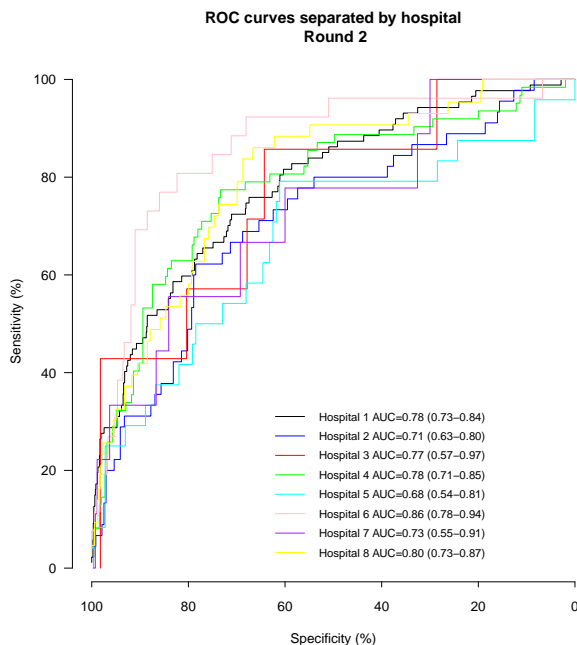


Figure 5.3: ROC curves of Computer-Aided Triaging for the task of distinguishing between examinations with lesions (benign and malignant) and examinations without lesions, applied to second screening-round data. The 95% confidence intervals are shown in the legend.

5.5.3 CAD

CAD, applied on all lesions in the dataset, classified 34 lesions (12 BI-RADS 3 and 22 BI-RADS 4) correctly as benign, and 75 as malignant (7 BI-RADS 3, 62 BI-RADS 4 and 6 BI-RADS 5) of which 20 (17 BI-RADS 4 and 3 BI-RADS 5) were malignant at histology. An increase in PPV was observed compared with unassisted radiological reading in both screening rounds, $p < 0.001$ (Table 5.3).

At the established operating point, the CAD shows no difference in results in the second screening round compared to the first round ($p=0.08$) (Table 5.3, Figure 5.4).

Table 5.3: Results of CAD and radiological reading in data from first and second screening round of DENSE. PPV=positive-predictive value, CAD= Computer-aided diagnosis

	First-round data	Second- round data	p-value
AUC	0.86 (0.81-0.90)	0.75(0.64-0.86)	p=0.08
Benign lesions classified as benign by CAD	41.0% (36.3%-45.8%, 176 of 429)	38.2% (28.1%-49.1%, 34 of 89)	p=0.62
PPV of CAD (BI-RADS 3-5)	23.6% (22.2%-25.1%; 77 of 326)	26.7%(23.6%-30.0%; 20 of 75)	p <0.001
PPV of radiological reading (BI-RADS 3-5)*	15.2% (14.4%-16.1%; 77 of 506)	18.4%(15.9%-21.1%; 20 of 109)	p <0.001

*Percentages can differ from earlier publication[15] because not all examination of screening round 2 were included.

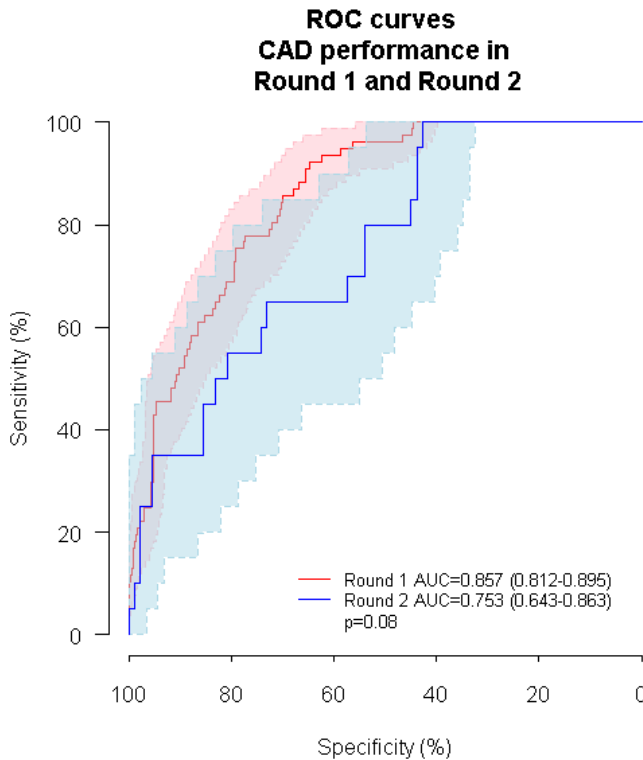


Figure 5.4: ROC curves of Computer-Aided Diagnosis for the task of distinguishing between benign and malignant lesions, applied to first and second screening round data. The shade regions represent the 95% confidence intervals.

5.5.4 Combination of computer-assisted triaging and computer-aided diagnosis

Combined CAT and CAD confirms potential to dismiss a subset of examinations and lesions for further assessment without missing any malignant disease (Figure 5.5). CAT would dismiss 950 / 2 901 (32.7%) examinations. Thirty-eight of 950 dismissed examinations (4.0%) contained one or more benign lesions. None were malignant.

In the remaining 1951 examinations, 265 examinations contained 285 lesions. CAD classified 132/285 (46.3%) of these lesions as benign. No malignant lesions were called benign. At best, the combination of CAT and CAD would yield 53/109 (48.6%) false-positive referrals to additional MRI screening or biopsy, compared to 89/109 (78.9%) for radiologists without computer assistance ($p=0.001$).

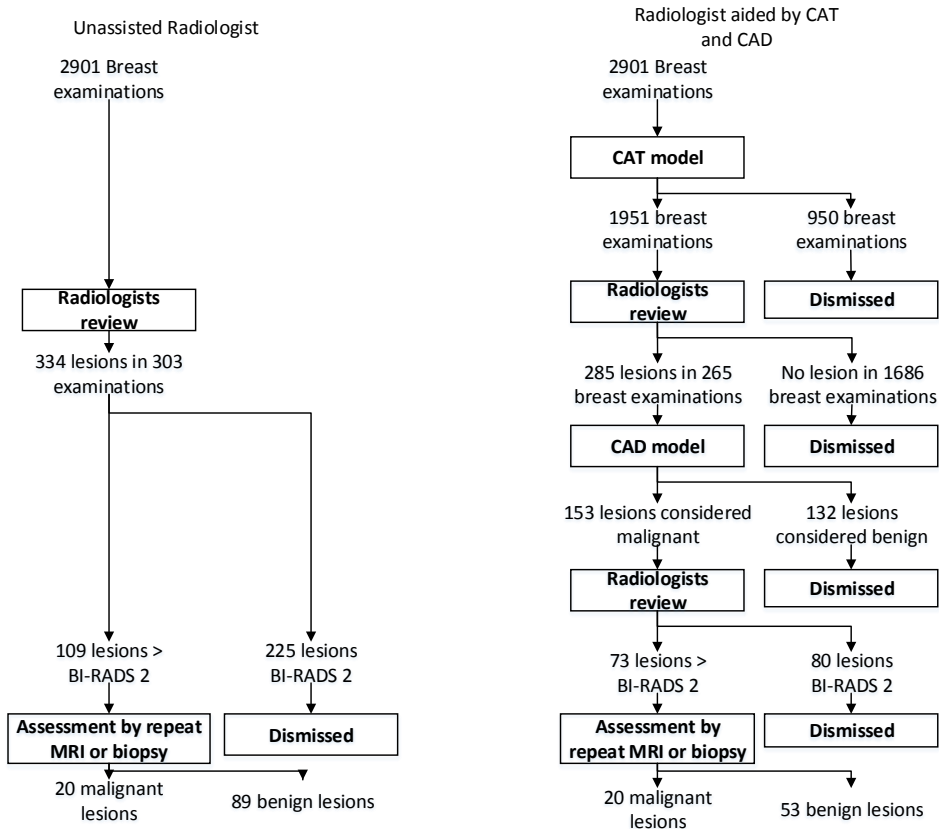


Figure 5.5: A comparison of the workload of the radiologist in terms of reading and workup with and without computerized analysis.
 NOTE: Two benign lesions that CAD would classify as probably malignant were dismissed by CAT, which explains why CAD classifies only 73 of the 75 as malignant as stated in the CAD results section.

5.5.5 Discussion

Adding MRI to breast cancer screening programs for women with extremely dense breasts will result in increased workload for radiologists. In order to reduce this workload and also the false-positive rate methods were developed for computer-assisted triaging (CAT) and computer-aided diagnosis (CAD)[7, 15]. Here we show that CAT and CAD developed on data from the first screening round of the DENSE trial, reproduce robustly in the second screening round: combined CAT and CAD has potential to reduce the workload of radiologists by 32.7% (950 of 2 901) by dismissing normal examinations without dismissing cancers. In the remaining scans considered to require reading, 132/285 (46.3%) lesions were

correctly identified as benign without missing cancers.

The combination of methods shows potential to reduce false-positive referrals by 40.4% (36 of 89).

Of the 89 benign lesions referred to biopsy or additional MRI after unassisted radiological reading, three benign lesions could be dismissed by CAT (3.4%), followed by 33 by CAD (36.0%), totaling 39.4%. To the best of our knowledge, no other groups than Verburg et al. and Dekker et al.[15, 21], reported on reduction of false-positive referrals in the MR screening of women with extremely dense breasts. For other breast MR indications, on the basis of smaller and more heterogeneous populations, reductions have been reported of 12 /24 (50.0%) with CAD[22], and 17/24 (70.8%) with proton MR-spectroscopy[23].

Although several studies reported on false-positive rates for computer-aided detection, none described the number of correctly identified normal breasts[6, 24–28].

Although the percentage of dismissed examinations without lesion did not differ between screening rounds ($p=0.70$), the AUC showed a difference ($p=0.001$). This indicates that differences are present in the appearance of lesions in round 1 and those in round 2. This may be caused by differences between a prevalent screening round and an incident screening round: lesions in the incident round became visible in a time span of two years where lesions in the prevalent round may have existed for a longer period of time.

Whereas the concept of automatically dismissing normal breast MRI examinations is attractive to reduce radiologist workload, and may be feasible from a technical perspective, challenges remain to clinical implementation. In current practice every screening image has to be interpreted by a trained physician. Before the required paradigm shift for implementation would be accepted by patients, clinicians and policy makers must address multiple issues like safety, accountability and quality [29]

This study also has limitations. Participants of the second round of the DENSE trial also participated in the first round; data were acquired in the same hospitals using the same MRI devices with identical sequences. The number of malignant lesions in the second screening round was limited. To further investigate robustness, methods should be tested on data acquired under more various conditions. Future studies could focus on application of presented methodology to other screening populations, such as women at high life-time risk of developing breast cancer. Also, the level of automation can be further increased; current operator

involvement was twofold: providing manual location of lesions for the CAD and manual identification of BIRADS-2 lesions after CAD. Future research will focus on automating these steps as well.

In conclusion, combining CAT and CAD has the potential to both reduce workload and reduce the number of biopsies without dismissing malignant breast disease.

References

- [1] M. F. Bakker et al., "Supplemental mri screening for women with extremely dense breast tissue," *New England Journal of Medicine*, vol. 381, no. 22, pp. 2091–2102, 2019.
- [2] E. Warner et al., "Systematic review: Using magnetic resonance imaging to screen women at high risk for breast cancer," *Annals of Internal Medicine*, vol. 148, no. 9, pp. 671–679, 2008.
- [3] S. Saadatmand et al., "Mri versus mammography for breast cancer screening in women with familial risk (famrisc): a multicentre, randomised, controlled trial," *The Lancet Oncology*, 2019.
- [4] G. L. Menezes et al., "Magnetic resonance imaging in breast cancer: A literature review and future perspectives," *World journal of clinical oncology*, vol. 5, no. 2, pp. 61–70, 2014.
- [5] G. Maicas et al., "Deep reinforcement learning for active breast lesion detection from dce-mri," *Medical Image Computing and Computer Assisted Intervention MICCAI 2017*, pp. 665–673, Springer International Publishing, 2017.
- [6] A. Vignati et al., "Performance of a fully automatic lesion detection system for breast dce-mri," *Journal of Magnetic Resonance Imaging*, vol. 34, no. 6, pp. 1341–1351, 2011.
- [7] E. Verburg et al., "Deep learning for automated triaging of 4 581 breast mris from the dense trial," *Radiology*, vol. Ahead of print, no. -, pp. -, 2021.
- [8] K. G. A. Gilhuijs, M. L. Giger, and U. Bick, "Computerized analysis of breast lesions in three dimensions using dynamic magnetic-resonance imaging," *Medical Physics*, vol. 25, no. 9, pp. 1647–1654, 1998.
- [9] E. Honda, R. Nakayama, H. Koyama, and A. Yamashita, "Computer-aided diagnosis scheme for distinguishing between benign and malignant masses in breast dce-mri," *Journal of digital imaging*, vol. 29, no. 3, pp. 388–393, 2016.
- [10] H. Rahbar and S. C. Partridge, "Multiparametric mr imaging of breast cancer," *Magnetic resonance imaging clinics of North America*, vol. 24, no. 1, pp. 223–238, 2016.
- [11] M. U. Dalmış et al., "Artificial intelligence–based classification of breast lesions imaged with a multiparametric breast mri protocol with ultrafast dce-mri, t2, and dwi," *Investigative Radiology*, vol. Publish Ahead of Print, 2019.
- [12] D. Truhn et al., "Radiomic versus convolutional neural networks analysis for classification of contrast-enhancing lesions at multiparametric breast mri," *Radiology*, vol. 290, no. 2, pp. 290–297, 2019.
- [13] C. Gallego-Ortiz and A. L. Martel, "Using quantitative features extracted from t2-weighted mri to improve breast mri computer-aided diagnosis (cad)," *PLOS ONE*, vol. 12, no. 11, p. e0187501, 2017.
- [14] N. Bhooshan et al., "Combined use of t2-weighted mri and t1-weighted dynamic contrast-enhanced mri in the automated analysis of breast lesions," *Magnetic resonance in medicine*, vol. 66, no. 2, pp. 555–564, 2011.

- [15] E. Verburg et al., "Computer-aided diagnosis in multiparametric magnetic resonance imaging screening of women with extremely dense breasts to reduce false-positive diagnoses," *Invest Radiol*, vol. 55, no. 7, pp. 438–444, 2020.
- [16] M. J. Emaus et al., "Mr imaging as an additional screening modality for the detection of breast cancer in women aged 50-75 years with extremely dense breasts: The dense trial study design," *Radiology*, vol. 277, no. 2, pp. 527–37, 2015.
- [17] S. G. A. Veenhuizen et al., "Supplemental breast mri for women with extremely dense breasts: Results of the second screening round of the dense trial," *Radiology*, vol. 299, no. 2, pp. 278–286, 2021.
- [18] E. Morris et al., "Acr bi-rads® atlas, breast imaging reporting and data system," Reston, VA: American College of Radiology, pp. 56–71, 2013.
- [19] T. Alderliesten et al., "Validation of semiautomatic measurement of the extent of breast tumors using contrast-enhanced magnetic resonance imaging," *Investigative Radiology*, vol. 42, no. 1, pp. 42–49, 2007.
- [20] Oncoline, "Borstkanker - algemeen," 2017.
- [21] B. M. d. Dekker et al., "Reducing false-positive screening mri rate in women with extremely dense breasts using prediction models based on data from the dense trial," *Radiology*, vol. 301, no. 2, pp. 283–292, 2021.
- [22] C. D. Lehman, S. Peacock, W. B. DeMartini, and X. Chen, "A new automated software system to evaluate breast mr examinations: improved specificity without decreased sensitivity," *American Journal of Roentgenology*, vol. 187, no. 1, pp. 51–56, 2006.
- [23] P. Clauser, M. Marcon, M. Dietzel, and P. A. T. Baltzer, "A new method to reduce false positive results in breast mri by evaluation of multiple spectral regions in proton mr-spectroscopy," *European Journal of Radiology*, vol. 92, pp. 51–57, 2017.
- [24] A. Gubern-Mérida et al., "Automated localization of breast cancer in dce-mri," *Medical Image Analysis*, vol. 20, no. 1, pp. 265–274, 2015.
- [25] M. U. Dalmış et al., "Fully automated detection of breast cancer in screening mri using convolutional neural networks," *Journal of medical imaging (Bellingham, Wash.)*, vol. 5, no. 1, pp. 014502–014502, 2018.
- [26] D. M. Renz et al., "Detection and classification of contrast-enhancing masses by a fully automatic computer-assisted diagnosis system for breast mri," *Journal of Magnetic Resonance Imaging*, vol. 35, no. 5, pp. 1077–1088, 2012.
- [27] Y.-C. Chang et al., "Computerized breast lesions detection using kinetic and morphologic analysis for dynamic contrast-enhanced mri," *Magnetic resonance imaging*, vol. 32, no. 5, pp. 514–522, 2014.
- [28] F. Ayatollahi, S. B. Shokouhi, R. M. Mann, and J. Teuwen, "Automatic breast lesion detection in ultrafast dce-mri using deep learning," *arXiv preprint arXiv:2102.03932*, 2021.

[29] B. N. Joe, "Ai to dismiss normal breast mri scans and reduce workload," *Radiology*, vol. 0, no. 0, p. 212247, 2021.

Chapter 6

Discussion



6.1 Conclusion

In this thesis we investigated computerized decision support for radiologic review of MRI screening examinations of women with extremely dense breasts. Results show promising potential to reduce the number of false positive findings and to prioritize the workflow.

First, chest wall segmentation in MRI of extremely dense breasts can be challenging because the contrast between muscle and glandular tissue is poor. Two methods were developed to accomplish this task: the knowledge based method (KBM) and the deep learning based method (DLM). These methods had similar performance: Dice similarity coefficient (DSC) metric of 0.982 and 0.984, respectively (**Chapter 2**).

Secondly, computer-aided triaging (CAT) was developed to dismiss normal breast examinations (i.e., examinations that show no lesions) from radiological review without missing malignant disease. The method showed potential to dismiss 39.7% of normal bilateral breast examinations (**Chapter 3**).

Thirdly, we addressed computer-Aided diagnosis (CAD) to distinguish between benign and malignant lesions in the first round of the dense trial. The presented method was able to correctly classify 41.5% (176 of 425) of benign BI-RADS 3 and BI-RADS 4 lesions as benign without missing malignant lesions (**Chapter 4**).

Fourthly, we combined CAT and CAD and validated the methods on the second screening round of the DENSE trial. This showed feasibility to dismiss 41.0% normal examinations without missing cancer, followed by correct identification of 46.3% (132 of 285) benign lesions by CAD. (**Chapter 5**).

6.2 Discussion

Although the DENSE trial showed that MRI screening in women with extremely dense breasts reduces the number of interval cancers, it also confirmed the limited specificity to detect breast cancer. Of the 4783 women screened, 454 women were referred to additional follow up, resulting in cancer detection in 79 women and reduction of the number of interval cancers from 4.9 to 0.8 per 1000 screenings[1].

In a first advice to the secretary of health of the Netherlands[2], the Dutch health council concluded that the benefits of MRI screening of women with extremely dense breasts (i.e., early detection of breast cancers), barely outweigh the disad-

vantages (i.e., false-positive outcomes, overdiagnosis, overtreatment, emotional and physical burden, workload and costs). They recommended to first investigate less expensive alternatives such as contrast enhanced mammography before investing in the infrastructure and staff for nationwide MRI screening. In addition, they expected that future developments in diagnostics, risk stratification and artificial intelligence may improve the cost-benefit balance of MRI screening. The Dutch government later accepted a resolution to make MRI screening available for women with extremely dense breast, pending supplemental investigation that compares contrast-enhanced mammography with MRI. Multiple approaches contribute to optimizing the cost-benefit of MRI, for example: methods to reduce the number of false positive referrals, methods to optimize the workflow, and methods to improve selection of the screening population. This thesis attempted to reduce the number of false-positive referrals, hence leading to less emotional and physical burden of screening participants and reduced clinical workload and costs. Both are important for a successful screening program. High emotional and physical burden will negatively affect the willingness of participants to take part in screening. Reduction of the number of additional MRI examinations and biopsies on benign lesions due to false-positive referral will also reduce the workload and costs. In addition, triaging of screening images allows the radiologists to prioritize their workload in order to optimize the workflow. Although MRI screening of women with extremely dense breasts already focuses on the additional risk caused by breast density, further risk stratification should improve the selection of the screening population to further increase the specificity and to reduce overdiagnosis and cost. First results of risk stratification in data acquired in the DENSE trial showed that the Tyrer-Cuzick 5-year[3] and BSCS 5-year[4] breast risk estimates have significantly higher cancer detection rates in the highest risk quartiles[5].

6.2.1 Opportunities for improvements

The methods developed in this thesis contain several aspects that can be further improved. In the following paragraphs, the opportunities for improvement for the segmentation to automatically delineate the chest wall in MRI images, the CAT to triage and the CAD to classify lesions are discussed. In the last paragraph we discuss the possibility for workflow automation of combined CAT and CAD.

Segmentation

Multiple automated segmentation methods demonstrated performances comparable to that of humans[6–8]. As presented in **Chapter 2** of this thesis, the segmentation of the chest wall can be successfully performed using knowledge-based and deep-learning based methods, however some room for improvement was found to be present. The performance of machine learning depends on the quality of and variation in the training data. To improve the methods, more training data that covers a larger variation of anatomical properties, MRI quality and MRI acquisitions are required. Hence, more complex methods, for example deep learning with more layers or knowledge-based methods with more input features, can be trained to improve, taking risk of overtraining into account.

In addition to the performance of the segmentation method, the methods could also be expanded to segment more tissue types. When all tissue types in the breast are segmented this can be used to harvest features for other purposes like triaging, risk stratification or diagnosis.

CAT

To our knowledge, the work presented in this thesis was the first to report on CAT for MRI screening images of women with extremely dense breasts. We were able to dismiss 39.7% (95%CI: 30.0%, 49.4%) of (bilateral) breast examinations without lesions without missing any malignant lesions. Although the absolute number of examinations that could be dismissed in nationwide screening of women with extremely dense breasts using CAT may be substantial, there is room for further improvement.

The MRI protocol in the DENSE trial was extensive. Nonetheless CAT only used 2D maximum intensity projections in three directions of the subtracted pre-contrast and post-contrast image series. These images typically show relatively high contrast uptake in lesions. Extending the training data with additional MRI sequences and using a more complex model to represent the 3D data is likely to improve the performance of the CAT.

CAD

Since onset of widespread clinical use of breast MRI, CAD has been investigated to assess breast lesions on MRI. Already in 1998 and 1999 authors were able to train models with decent performances on small cohorts of symptomatic women.

Later, multiple authors developed additional CAD methods using several techniques, all resulting in comparable results (Table 6.1)

Table 6.1: Overview of performances of CAD methods developed between 1998 and 2020. Metric Area under the ROC curve (AUC) reflects the performance of the model.

Author	Year	AUC	Benign	Malignant	Multicenter	Method
Gilhuijs[9]	1998	0,96	13	15	No	Stepwise multiple regression
Penn[10]	1999	0,86	20	32	No	Logistic Regression
Gilhuijs[11]	2002	0,95	40	40	No	Logistic Regression
Chen[12]	2004	0,86	44	77	No	Linear Discriminant Analysis
Chen[13]	2006	0,85	44	77	No	Fuzzy C-Means
Newell[14]	2010	0,86	28	88	No	Artificial Neural Network
Bhooshan[15]	2011	0,85	86	110	No	Artificial Neural Network
Hoffmann[16]	2013	0,87	23	61	No	Quadratic discriminant analysis
Wang[17]	2014	0,74	31	131	No	Random forest
Gallego-Ortiz[18]	2014	0,9	100	143	No	Random forest
Razavi[19]	2016	0,9	38	68	No	Random forest
Gallego-Ortiz[20]	2017	0,83	382	245	No	Boosted classification tree
Ji[21]	2019	0,89	496	1483	No	Support Vector Machine
Dalmis[22]	2019	0,85	149	368	No	Convolutional Neural Network
Truhn[23]	2019	0,88	507	787	No	Convolutional Neural Network
Truhn[23]	2019	0,81	507	787	No	Lasso regression
Verburg[24] (this thesis)	2020	0,85	429	77	Yes	Ridge regression

In the past 22 years, the size of data sets grew and the models became more complex, however the performance of the models did not improve noticeably. Although the results are difficult to compare because different data sets were used, the question arises whether we have reached the limits of the performance of CAD models in breast MRI. We believe that room for improvement could be present in capturing data complexity (1), data curation (2) and bias prevention (3).

Capturing data complexity

Distinguishing benign from malignant lesions based on MRI examinations is a complex challenge. More than 20 years of development did not result in a model with performance matching that of pathology. It would appear that we may not be able to fully extract the required information from the MRI examinations for this purpose. A possible explanation is that although MRI has sub-millimeter resolution, the amount of information in one voxel is still substantially smaller than the amount of biological information in the cells that span one voxel dimension. Additional detailed information is not accessible for MRI. New MRI techniques to

image other elements like fluorine, phosphor, sodium and carbon, may contribute to a solution.

Another part which may be considered too simplistic is the ground truth. In general, and in this thesis, we assume that there are two types of lesions, malignant and benign, however in reality this is more complex. Malignant is defined as tending to infiltrate, metastasize, and terminate fatally, however some lesions defined as malignant will never progress to invasive disease[25]. A more complex ground truth will make it possible to distinguish more types of lesions and possibly give new insights in differentiating features.

Data curation

Important sources of variation are the MRI-devices from multiple vendors and qualitative nature of MRI data, where voxel values often do not contain more information than image intensity. The large amount of variation present in MRI data obtained under different circumstances makes it complex to create a single model for lesion classification. Although the MRI protocol of the DENSE trial is standardized, some variation was present between the datasets acquired in the different hospitals. The T2-weighted sequence was optional and therefore not always present, T1- and T2-weighted images were acquired with and without fat suppression and DWI sequences were acquired with 2, 3 or 4 b-values. Moreover, some differences occurred in image dimensions and timing of dynamic series.

In this thesis we used methods for data harmonization and standardization. This resulted in a CAT that performed well on data acquired in a hospital of which no data was present in the training data. However, there was still a difference in performance on data acquired in different hospitals. For CAD the differences between hospitals were not visible because the data was pooled and randomly assigned to one of the 10 folds used during cross validation. Improving the harmonization, and thereby further reducing the variation in the data, will likely improve a model's performance. Another, less feasible solution, is to limit the variation in the data at the source.

Harmonized data will be present when identical MRI devices with identical sequences are used. However this may not be feasible because of the preference of hospitals and because it will limit the development of new MRI applications and techniques in screening.

MR image intensities do not have physical units. In addition, inhomogeneities in the magnetic field of an MRI device affect the voxel intensity. Data normalization

methods were developed to reduce this source of variation[26]. In addition, new quantitative techniques are being developed to prevent the bias at the source[27–29]. Using these methods it may be possible to replace the reconstructed qualitative voxel intensities with quantitative values of the proton density, the longitudinal (T_1) and transverse (T_2) relaxation rates. Less variation will be present in quantitative MRI images, thus reducing the need for data harmonization.

Bias prevention

Reported CAD models (Table 6.1) have been trained using relatively small datasets, which may not be representative for the general population of subjects of interest. The CAD developed in this thesis was trained to classify lesions in asymptomatic women with extremely dense breasts. Although the database of lesions was large ($n=506$), it did not cover the complete spectrum of possible lesion types. To prevent bias, larger data sets which contain data of more lesions types are required. Preferably, training data obtained from multiple hospitals using multiple vendor MRI devices and larger heterogeneity in participant population should be used to improve the performance.

Workflow automation

In addition to performance, also the usability of the presented methods in this thesis can be improved. This thesis shows that the combination of CAT and CAD has complimentary value. Although the larger part of the method was automated, still some manual steps were required. The lesion seed point was placed manually. Future developments of the presented CAT and CAD should focus on fully automating the workflow. The results of the chestwall segmentation (**Chapter 2**) should be used to select a region of interest which can be input for the CAT (**Chapter 3**), and the output of the CAT should be used for lesion localization required for the CAD (**Chapter 4**).

Seed point placement can be automated with use of the SHAP output from the CAT. SHAP highlights the areas in the breast responsible for the output of the CAT. In most cases the lesion, when present, is indicated by SHAP and this information can be used to replace the manual input for the lesion seed point. Fully automated flow can assist the radiologist in two ways. First it can prioritize the workflow, and secondly it has the potential to reduce false positive referrals. It is, however, important to investigate the ratio of false positive lesions localized based

on triaging. Other studies showed 4 to 6 false positives at sensitivity of 0.89-0.94 for lesions detection[30–32].

6.2.2 Future perspective

In this thesis we showed the potential of computerized analysis applied to screening MRI of women with extremely dense breasts by detecting examinations without lesions and to reduce false positive referrals. Before these methods can be implemented in clinical practice they have to be extensively validated because errors by the models can result in missing a malignant lesion. In other words, the sensitivity of the models for detection of malignant lesions should at least be as good as the sensitivity of a well-trained radiologist.

First, the interaction of the computerized models with radiologists needs to be investigated in the clinical workflow. In this thesis we demonstrated the proof of concept of autonomous analysis, but did not investigate the interaction between the software and the radiologists.

Secondly, the stability of the methods could be investigated by testing on consecutively included data of women scanned for breast cancer in additional medical centers. Newly developed software methods and vendor agnostic artificial intelligence platforms could make it possible to apply computerized methods on medical images in the clinical workflow. Another important aspect to consider is the ethical side. It is theoretically possible that the models do miss a malignant lesion by dismissing an examination with malignant lesions or classifying a malignant lesion as benign. In this thesis we present results when 100% of the malignant lesions are triaged to radiological review. The thresholds of what is acceptable are hard to set (100% or less), in our opinion we should always strive to detect all malignant lesions.

Disadvantages of MRI screening are the acquisition time and the workload for radiologist caused by the large amount of acquired data. To reduce acquisition time and limit the amount of acquired data, studies have investigated abbreviated and high-temporal resolution MRI protocols for breast cancer diagnosis and screening. In abbreviated protocols the number of acquisitions is reduced to typically one pre-contrast and one post-contrast T1-weighted imaged, sometimes complemented with a T2-weighted image[33, 34].

In this thesis we show that segmentation of the chest wall, CAD to distinguish benign lesions from malignant lesions and computer aided triaging to dismiss

breasts without lesion all could be used on data acquired using an abbreviated protocol. In addition, the CAD (**Chapter 4**) also demonstrated additional value of a full parametric MRI protocol over an abbreviated protocol: The specificity at 100% sensitivity for malignant lesions decreased from 41.5% to 26.2% when only features from an abbreviated protocol were used. Hence, the larger acquisition time and workload of the radiologist in a full multiparametric protocol may result in less false-positive referrals to biopsy or additional MRI screening without loss of sensitivity for malignant lesions, which would also save time and resources.

References

- [1] M. F. Bakker et al., "Supplemental mri screening for women with extremely dense breast tissue," *New England Journal of Medicine*, vol. 381, no. 22, pp. 2091–2102, 2019.
- [2] Gezondheidsraad, 2020.
- [3] J. Tyrer, S. W. Duffy, and J. Cuzick, "A breast cancer prediction model incorporating familial and personal risk factors," *Stat Med*, vol. 23, no. 7, pp. 1111–30, 2004.
- [4] J. A. Tice et al., "Breast density and benign breast disease: Risk assessment to identify women at high risk of breast cancer," *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, vol. 33, no. 28, pp. 3137–3143, 2015.
- [5] S. de Lange et al., "Screening performance of supplemental mri in women with extremely dense breasts stratified by breast cancer risk," *X*, vol. X, no. X, pp. 999–999, 2020.
- [6] R. Meier et al., "Clinical evaluation of a fully-automatic segmentation method for longitudinal brain tumor volumetry," *Sci Rep*, vol. 6, p. 23376, 2016.
- [7] J. Wong et al., "Comparing deep learning-based auto-segmentation of organs at risk and clinical target volumes to expert inter-observer variability in radiotherapy planning," *Radiotherapy and Oncology*, vol. 144, pp. 152–158, 2020.
- [8] J. Ma et al., "Abdomenct-1k: Is abdominal organ segmentation a solved problem?," *arXiv preprint arXiv:2010.14808*, 2020.
- [9] K. G. A. Gilhuijs, M. L. Giger, and U. Bick, "Computerized analysis of breast lesions in three dimensions using dynamic magnetic-resonance imaging," *Medical Physics*, vol. 25, no. 9, pp. 1647–1654, 1998.
- [10] A. I. Penn, L. Bolinger, M. D. Schnall, and M. H. Loew, "Discrimination of mr images of breast masses with fractal-interpolation function models," *Academic Radiology*, vol. 6, no. 3, pp. 156–163, 1999.
- [11] K. G. A. Gilhuijs et al., "Breast mr imaging in women at increased lifetime risk of breast cancer: Clinical system for computerized assessment of breast lesions—initial results," *Radiology*, vol. 225, no. 3, pp. 907–916, 2002.
- [12] W. Chen, M. L. Giger, L. Lan, and U. Bick, "Computerized interpretation of breast mri: investigation of enhancement-variance dynamics," *Med Phys*, vol. 31, no. 5, pp. 1076–82, 2004.
- [13] W. Chen, M. L. Giger, U. Bick, and G. M. Newstead, "Automatic identification and classification of characteristic kinetic curves of breast lesions on dce-mri," *Med Phys*, vol. 33, no. 8, pp. 2878–87, 2006.
- [14] D. Newell et al., "Selection of diagnostic features on breast mri to differentiate between malignant and benign lesions using computer-aided diagnosis: differences in lesions presenting as mass and non-mass-like enhancement," *European Radiology*, vol. 20, no. 4, pp. 771–781, 2010.

- [15] N. Bhooshan et al., "Combined use of t2-weighted mri and t1-weighted dynamic contrast-enhanced mri in the automated analysis of breast lesions," *Magnetic resonance in medicine*, vol. 66, no. 2, pp. 555–564, 2011.
- [16] S. Hoffmann et al., "Automated analysis of non-mass-enhancing lesions in breast mri based on morphological, kinetic, and spatio-temporal moments and joint segmentation-motion compensation technique," *EURASIP Journal on Advances in Signal Processing*, vol. 2013, no. 1, p. 172, 2013.
- [17] L. Wang et al., "A robust and extendable framework towards fully automated diagnosis of non-mass lesions in breast dce-mri," in *2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI)*, pp. 129–132, IEEE, 2014.
- [18] C. Gallego-Ortiz and A. L. Martel, "Classification of breast lesions presenting as mass and non-mass lesions," in *Medical Imaging 2014: Computer-Aided Diagnosis*, vol. 9035, p. 90351Z, International Society for Optics and Photonics, 2014.
- [19] M. Razavi et al., "Novel morphological features for non-mass-like breast lesion classification on dce-mri," *Machine Learning in Medical Imaging*, pp. 305–312, Springer International Publishing, 2016.
- [20] C. Gallego-Ortiz and A. L. Martel, "Using quantitative features extracted from t2-weighted mri to improve breast mri computer-aided diagnosis (cad)," *PLOS ONE*, vol. 12, no. 11, p. e0187501, 2017.
- [21] Y. Ji et al., "Independent validation of machine learning in diagnosing breast cancer on magnetic resonance imaging within a single institution," *Cancer Imaging*, vol. 19, no. 1, p. 64, 2019.
- [22] M. U. Dalmış et al., "Artificial intelligence–based classification of breast lesions imaged with a multiparametric breast mri protocol with ultrafast dce-mri, t2, and dwi," *Investigative Radiology*, vol. Publish Ahead of Print, 2019.
- [23] D. Truhn et al., "Radiomic versus convolutional neural networks analysis for classification of contrast-enhancing lesions at multiparametric breast mri," *Radiology*, vol. 290, no. 2, pp. 290–297, 2019.
- [24] E. Verburg et al., "Computer-aided diagnosis in multiparametric magnetic resonance imaging screening of women with extremely dense breasts to reduce false-positive diagnoses," *Invest Radiol*, vol. 55, no. 7, pp. 438–444, 2020.
- [25] L. J. Grimm et al., "Surgical upstaging rates for vacuum assisted biopsy proven dcis: Implications for active surveillance trials," *Annals of Surgical Oncology*, vol. 24, no. 12, pp. 3534–3540, 2017.
- [26] N. J. Tustison et al., "N4itk: improved n3 bias correction," *IEEE Trans Med Imaging*, vol. 29, no. 6, pp. 1310–20, 2010.
- [27] H. Margaret Cheng, N. Stikov, N. R. Ghugre, and G. A. Wright, "Practical medical applications of quantitative mr relaxometry," *Journal of Magnetic Resonance Imaging*, vol. 36, no. 4, pp. 805–824, 2012.

- [28] D. Ma et al., "Magnetic resonance fingerprinting," *Nature*, vol. 495, no. 7440, pp. 187–192, 2013.
- [29] A. Sbrizzi et al., "Fast quantitative mri as a nonlinear tomography problem," *Magnetic Resonance Imaging*, vol. 46, pp. 56–63, 2018.
- [30] S. B. Shokouhi, A. Fooladivanda, and N. Ahmadinejad, "Computer-aided detection of breast lesions in dce-mri using region growing based on fuzzy c-means clustering and vesselness filter," *EURASIP Journal on Advances in Signal Processing*, vol. 2017, no. 1, p. 39, 2017.
- [31] A. Gubern-Mérida et al., "Automated localization of breast cancer in dce-mri," *Medical Image Analysis*, vol. 20, no. 1, pp. 265–274, 2015.
- [32] Y.-C. Chang et al., "Computerized breast lesions detection using kinetic and morphologic analysis for dynamic contrast-enhanced mri," *Magnetic resonance imaging*, vol. 32, no. 5, pp. 514–522, 2014.
- [33] C. K. Kuhl et al., "Abbreviated breast magnetic resonance imaging (mri): first postcontrast subtracted images and maximum-intensity projection—a novel approach to breast cancer screening with mri," *J Clin Oncol*, vol. 32, p. 2304, 2014.
- [34] D. Leithner et al., "Abbreviated mri of the breast: does it provide value?," *Journal of Magnetic Resonance Imaging*, vol. 49, no. 7, pp. e85–e100, 2019.

Chapter 7

Appendix



English summary

Women with extremely dense breasts (Breast Imaging Reporting and Data System [BI-RADS] class D), that is, breasts containing a large amount of fibroglandular tissue, have a 2 to 6 times higher risk of developing breast cancer than women with very fatty breasts. Moreover, these cancers are harder to detect on mammography due to the low contrast between fibroglandular tissue and tumor tissue and overlapping tissue. Additional MRI screening for these women has proven to detect additional cancers and reduce the number of interval cancers at the cost of false positive referrals and increased workload.

In this thesis the data acquired during the first and second screening round of the DENSE trial was used to develop and validate two computerized methods to reduce the number of false positive referrals and to triage the acquired MRI screening examinations.

The DENSE trial MRI protocol consisted of 5 different MRI sequences: T1-weighted imaging without fat suppression, diffusion-weighted imaging, T1-weighted contrast-enhanced images at high spatial resolution, T1-weighted contrast-enhanced images at high temporal resolution, and T2-weighted imaging.

A typical first step in computerized methods often is the definition of the region of interest, the breast tissue. Breast tissue is surrounded by skin and the chest wall, especially the latter is challenging to delineate in patients with dense breasts because the contrast between muscle and glandular tissue is poor. The aim of **Chapter 2** was to develop and compare a knowledge-based (KBM) and a deep learning-based approach (DLM) for segmentation of the chest wall in MR images of extremely dense breasts.

The KBM used shape, location, and gradient features, and the deep learning-based method (DLM) used a dilated convolution neural network. A data set of 115 T1-weighted MR images was randomly selected from MR images of women with extremely dense breasts participating in the DENSE trial. Manual segmentations of the chest wall, acquired under supervision of an experienced breast radiologist, were available for all data sets and used as ground truth. Both methods were optimized using the same randomly selected 36 MRI data sets. In the

remaining 79 data sets, the results of both segmentation methods were qualitatively evaluated. A radiologist reviewed the segmentation results of both methods in all transversal images ($n = 14\ 141$). It was determined whether the result would impact the ability to accurately determine the volume of fibroglandular and fatty tissue and whether breast regions that might harbor lesions were masked by the segmentation result. In addition, all segmentations were quantitatively assessed using the Dice similarity coefficient (DSC) and Hausdorff distance (HD), false positive fraction (FPF), and false negative fraction (FNF) metrics.

According to the radiologist's evaluation, the DLM had a significantly higher success rate than the KBM (81.6% vs 78.4%, $p < 0.01$). The success rate was further improved to 92.1% by combining both methods. Similarly, the DLM had significantly lower values for FNF (0.003 ± 0.003 vs 0.009 ± 0.011 , $p < 0.01$) and HD95 (2.58 ± 1.78 mm vs 3.37 ± 2.11 , $p < 0.01$). However, the KBM resulted in a significantly lower FPF than the DLM (0.018 ± 0.009 vs 0.030 ± 0.009 , $p < 0.01$). There was no significant difference between the KBM and DLM in terms of DSC (0.982 ± 0.006 vs 0.984 ± 0.008 , $p = 0.08$) or HD (24.14 ± 20.69 mm vs 12.81 ± 27.28 mm, $p = 0.05$).

Both optimized KBM and DLM showed good results to segment the pectoral muscle in women with dense breasts. Qualitatively assessed, the DLM was the most robust method. A quantitative comparison did not indicate, however, a preference for one method over the other.

To reduce the workload and to prioritize the work of breast MR radiologists, in **Chapter 3** automated triaging (CAT) is pursued that dismisses the highest number of examinations without lesions while still identifying all examinations with malignant disease.

A convolutional neural network (CNN) was developed to distinguish between breasts with lesions ($n=838$ BI-RADS 2 through 5, of which 77 malignant), and breasts without lesions ($n= 8\ 324$). The CNN was trained and validated using eight-fold internal-external validation. Here, the CNN is trained on seven hospitals and tested on the eighth hospital, alternating such that each hospital is tested once on independent data. The performance was assessed using receiver-operating characteristic (ROC) analysis. At 100% sensitivity for malignant disease, we es-

estimated the fraction of examinations that would be dismissed from radiological review.

At the operating point that detects all cancers, the DL model considered 90.7% (95%CI: 86.7%,94.7%) of the MRIs with lesions to be non-normal (i.e., contained BI-RADS 2, 3, 4 or 5 lesions), and triaged them to radiological review. The DL model dismissed 39.7% (95%CI: 30.0%, 49.4%) of the MRIs without lesions.

In **Chapter 4** a multiparametric machine-learning method was developed to predict, without deterioration of sensitivity, which of the BI-RADS 3- and BI-RADS 4-scored lesions are actually benign and could be prevented from being recalled.

In total, 506 lesions in 436 women were given a BI-RADS 3, 4 or 5 score. Of these lesions, 429 were benign and 77 were malignant on histologic examination. The CAD consists of 2 stages: feature extraction and lesion classification. Two groups of features were extracted: the first based on all multiparametric MRI sequences, the second based only on sequences that are typically used in abbreviated MRI protocols. In the first group, 49 features were used as candidate predictors: 46 were automatically calculated from the MRI scans, supplemented with 3 clinical features (age, body mass index, and BI-RADS score). In the second group, 36 image features and the same 3 clinical features were used. Each group was considered separately in a machine-learning model to differentiate between benign and malignant lesions. We developed a Ridge regression model using 10-fold cross validation.

Of the total number of BI-RADS 3 and BI-RADS 4 lesions referred to additional MRI or biopsy, 425/487 (87.3%) were false-positive. The full multiparametric model classified 176 (41.5%) and the abbreviated-protocol model classified 111 (26.2%) of the 425 false-positive BI-RADS 3- and BI-RADS 4-scored lesions as benign without missing a malignant lesion. If the full multiparametric CAD had been used to aid in referral, recall for biopsy or repeat MRI could have been reduced from 425/487 (87.3%) to 311/487 (63.9%) lesions. For the abbreviated protocol, it could have been 376/487 (77.2%).

In **Chapter 5**, the CAT of **Chapter 3** and CAD of **Chapter 4** were integrated to one flow where first the images were triaged and lesion triaged to radiological review

were classified by the CAD. The combined method was trained on the data of the first round and applied on second round data.

We estimate the impact of combined CAT and CAD in data from the second screening round of the DENSE trial and compare it with conventional radiological reading. We included all 2 901 breast MRI scans from the second round, obtained from eight hospitals in the Netherlands. The reproducibility of CAT and CAD were assessed by comparing results from the first and second screening rounds.

In second screening round examinations CAT would dismiss 950/2 901 (32.7%) examinations. Subsequent CAD would classify 132/285 (46.3%) of the non-dismissed lesions as benign without misclassifying any malignant lesion as benign. Combined CAT and CAD would result in significantly less false-positive lesions, 53/334 (15.9%) vs 89/334 (26.6%) ($p=0.001$) compared with the results from radiological reading.

Nederlandse samenvatting (Dutch summary)

Vrouwen met zeer dicht borstweefsel (Breast Imaging Reporting and Data System [BI-RADS] klasse D), dwz. borsten die veel klieren en weinig vet bevatten, hebben 2 tot 6 keer zoveel kans op het krijgen van borstkanker dan vrouwen met borsten die vooral uit vetweefsel bestaan. Daarnaast is kanker moeilijker te ontdekken in een mammogram omdat er weinig verschil is in contrast tussen klierweefsel en tumorweefsel. Het is bewezen dat aanvullende MRI screening, voor vrouwen met zeer dicht borstweefsel, extra kankers opspoort en het aantal gevonden kankers tussen screeningsronden verminderd, maar het zorgt ook voor meer fout positieve doorverwijzingen en meer werkdruk in het ziekenhuis.

In dit proefschrift is de data verzameld van de eerste en tweede ronde van de DENSE trial. Deze data is gebruikt om geautomatiseerde methoden te ontwikkelen die het aantal fout positieve verwijzingen verminderen en voor de triage van de MRI onderzoeken naar radiologen. Het DENSE trial MRI protocol bestond uit 5 verschillende MRI opnamen: T1-gewogen beelden zonder vet-onderdrukking, diffusie gewogen, T1-gewogen beelden met een hoge spatiele resolutie en contrast middel, T1-gewogen beelden met een hoge temporele resolutie en contrast middel en T2-gewogen beelden.

De eerste stap in geautomatiseerde methoden is vaak het identificeren van het gebied van interesse, het borstweefsel. In MR afbeeldingen wordt de borst omringd door huid en de borstwand. Vooral de borstwand is uitdagend om te lokaliseren in MR afbeeldingen van borsten met zeer dicht borstweefsel omdat er bijna geen contrastverschil is tussen spierweefsel en klierweefsel. Het doel in **Hoofdstuk 2** was het ontwikkelen en vergelijken van een op kennis gebaseerde methode (KBM) en een op deep-learning gebaseerde methode (DLM) voor het segmenteren van de borstwand in MR afbeeldingen van borsten met zeer dicht borstweefsel.

De KBM maakt gebruik van vorm, locatie en gradiënt eigenschappen, en de DLM gebruikt een convolutioneel neuraal netwerk. Een dataset van 115 T1-gewogen MR beelden van borsten met zeer dicht borstweefsel werd willekeurig geselecteerd uit de MR beelden van de DENSE trial. Een handmatige segmentatie van de borstwand die was gemaakt onder supervisie van een ervaren radioloog was beschikbaar voor elke dataset. De handmatige segmentatie werd

gebruikt om de resultaten van de automatische segmentaties mee te vergelijken. Een willekeurig geselecteerde dataset van 36 MR beelden werd gebruikt voor het optimaliseren van beide methoden. De resultaten van beide werden geëvalueerd in de overige 79 datasets. Segmentatieresultaten van beide methoden werden beoordeeld door een radioloog in alle transversale afbeeldingen ($n = 14\ 141$). Van elke segmentatie werd de invloed op het uitrekenen van klierweefselvolume en vetweefselvolume bepaald. Daarnaast werd beoordeeld of de segmentatie de detectie van borstlaesies zou kunnen beïnvloeden. Ook werden alle segmentaties kwantitatief beoordeeld met behulp van de Dice coëfficiënt (DSC), Hausdorff afstand (HD) en 95ste percentiel van de Hausdorff afstand (HD95), de fout-positief fractie (FPF) en de fout-negatief fractie (FNF).

Volgens de evaluatie van de radioloog had de DLM een significant hogere succesratio dan de KBM (81.6% tegenover 78.4%, $p < 0.01$). De succesratio verbetert tot 92.1% als je beide methoden samen gebruikt. De DLM had significant lagere waarden voor FNF (0.003 ± 0.003 tegenover 0.009 ± 0.011 , $p < 0.01$) en HD95 (2.58 ± 1.78 mm tegenover 3.37 ± 2.11 , $p < 0.01$). De KBM daarentegen, behaalde een significant lagere FPF dan de DLM (0.018 ± 0.009 tegenover 0.030 ± 0.009 , $p < 0.01$). Er was geen significant verschil in de vergelijking tussen de KBM en DLM in termen van DSC (0.982 ± 0.006 tegenover 0.984 ± 0.008 , $p = 0.08$) of HD (24.14 ± 20.69 mm tegenover 12.81 ± 27.28 mm, $p = 0.05$).

Beide geoptimaliseerde methoden resulteerden in goede segmentaties van de borstwand in MR beelden van vrouwen met zeer dicht borstweefsel. Kwalitatief is de DLM de meest robuuste methode, kwantitatief was geen van beide methodes beter dan de andere.

Om de werkdruk te verlagen en het werk van de radioloog te prioriteren is automatische triage (CAT) het doel van **Hoofdstuk 3**. Hierbij worden zoveel mogelijk MRI onderzoeken zonder laesies afgevangen en alle onderzoeken met kwaadaardige tumoren doorgestuurd.

Een convolutioneel neurale netwerk (CNN) werd ontwikkeld dat onderscheid kan maken tussen borsten met laesies ($n=838$ BI-RADS 2 tot en met 5, waarvan 77 kwaadaardig), en borsten zonder laesies ($n= 8\ 324$). Het CNN werd getraind en gevalideerd met behulp van een achtvoudige interne-externe validatie. Hier-

bij werd het CNN getraind op data uit zeven ziekenhuizen en getest op de data van het achtste ziekenhuis. Dit werd herhaald zodat data van elk ziekenhuis 1 keer werd gebruikt als test data. De methode werd beoordeeld door gebruik te maken van de receiver-operating karakteristiek (ROC) analyse. De fractie van afgevangen MRI onderzoeken werd ingesteld op het drempelwaarde waarbij de sensitiviteit voor het doorsturen van kwaadaardige tumoren de radioloog 100% is. Op de drempelwaarde waarbij alle kankers werden gedetecteerd, werd 90.7% (95%CI: 86.7%,94.7%) van de MRI-onderzoeken met laesies als niet normaal geclassificeerd (dus een BI-RADS 2, 3, 4 of 5 laesie), en derhalve doorgestuurd naar beoordeling door een radioloog. Het CNN ving 39.7% (95%CI: 30.0%, 49.4%) van de MRI onderzoeken zonder laesies af.

In **Hoofdstuk 4** werd een Computer Aided Diagnosis methode (CAD) ontwikkeld om te voorspellen welke BI-RADS 3- en BI-RADS 4-gescoorde laesies goedaardig zijn en in principe dus niet doorgestuurd hoeven worden voor vervolgonderzoek.

In totaal kregen 506 laesies in 436 vrouwen een BI-RADS 3, 4 of 5 score. Van deze laesies waren er 429 goedaardig en 77 kwaadaardig volgens histologische beoordeling. De CAD bestaat uit 2 stadia: het verkrijgen van bruikbare waarden (features) voor het model en laesie classificatie. Er werden twee groepen features verzameld, de eerste gebaseerd op alle multiparametrische MR beelden, de tweede gebaseerd op alleen de beelden die onderdeel zijn van een verkort MRI protocol. In de eerste groep werden 49 features gebruikt als mogelijke voorspeller: 46 daarvan werden automatisch bepaald uit de MR beelden en de andere 3 waren klinische features (leeftijd, Body-mass index [BMI] en BI-RADS score). In de tweede groep werden 36 features bepaald uit MR beelden, ook deze werden aangevuld met bovengenoemde 3 klinische features. Beide groepen werden onafhankelijk van elkaar gebruikt in een machine-learning model om het onderscheid te maken tussen goedaardige en kwaadaardige laesies. Hiervoor werd een Ridge regressie model ontwikkeld in combinatie met 10-voudige kruisvalidatie.

Van het totaal aantal BI-RADS 3- en BI-RADS 4-gescoorde laesies dat werd doorgestuurd voor een extra MRI of biopsie waren er 425/487 (87.3%) fout positief. Het volledige multiparametrische model classificeerde 176 (41.5%) laesies als goedaardig en het verkorte-protocol model classificeerde 111 (26.2%) van de 425 fout-positief BI-RADS 3- en BI-RADS 4-gescoorde laesies als goedaardig

zonder een kwaadaardige laesie te missen.

In **Hoofdstuk 5** werden de CAT van **Hoofdstuk 3** en de CAD van **Hoofdstuk 4** geïntegreerd in één werkstroom waarbij eerst de beelden werden geprioriteerd met triage en daarna de doorgestuurde laesies werden beoordeeld door de CAD. Beide methoden werden getraind op data uit de eerste screening ronde en toegepast op data uit de tweede screening ronde van de DENSE trial.

De impact van de gecombineerde toepassing van CAT en CAD op de data van de tweede screening ronde van de DENSE trial werd bepaald door het te vergelijken met conventionele beoordeling door een radioloog. In totaal werden 2901 borst MRI onderzoeken uit de 8 deelnemende ziekenhuizen geïnccludeerd. Daarnaast werd de reproduceerbaarheid van CAT en CAD methode beoordeeld door de resultaten van beide methoden te vergelijken tussen eerste en tweede ronde van de DENSE trial.

CAT ving in de tweede screening ronde 950/2 901 (32.7%) MRI onderzoeken af. CAD was daarna in staat om 132/285 (46.3%) van de doorgestuurde laesies te classificeren als goedaardig zonder een enkele kwaadaardige laesie als goedaardig te classificeren. Gecombineerd gebruik van CAT en CAD zou kunnen resulteren in significant minder fout-positief doorgestuurde laesies, 53/334 (15.9%) tegenover 89/334 (26.6%) ($p=0.001$) van de beoordeling door de radioloog.

Acknowledgments

The authors acknowledge the study participants for their contributions. This study is financially supported by KWF, grant number UU-2014-7151 and used data acquired during the DENSE trial. The DENSE trial was supported by the regional screening organisations, Volpara Solutions, the Dutch Expert Centre for Screening, and the National Institute for Public Health and the Environment. The DENSE trial is financially supported by the University Medical Center Utrecht (Project number: UMCU DENSE), the Netherlands Organization for Health Research and Development (ZonMw, Project numbers: ZONMW-200320002-UMCU and ZonMW Preventie 50-53125-98-014), the Dutch Cancer Society (KWF Kankerbestrijding, Project numbers: DCS-UU-2009-4348, UU-2014-6859 and UU2014-7151), the Dutch Pink Ribbon/A Sister's Hope (Project number: Pink Ribbon-10074), Bayer AG Pharmaceuticals, Radiology (Project number: BSP-DENSE), and Stichting Kankerpreventie Midden-West. The authors thank the registration team of the Netherlands Comprehensive Cancer Organization (IKNL) for the collection of data for the Netherlands Cancer Registry. The authors have no relevant conflicts of interest to disclose.

Dankwoord

Dit proefschrift was nooit tot stand gekomen zonder de bijdragen en hulp van velen. Een aantal mensen wil ik op deze plek in het bijzonder bedanken.

Eerst wil ik alle deelnemers van de DENSE-studie bedanken, die door deel te nemen aan de aanvullende MRI-screening aan de basis stonden van dit onderzoek.

Geachte prof. dr. ir. M.A. Viergever, beste Max, bedankt voor je interesse en je scherpe blik.

Geachte Dr. Gilhuijs, beste Kenneth, taal is zeg maar niet echt mijn ding. Bedankt voor je oneindige geduld tijdens de reviews van mijn schrijfsels. Het kan niet anders dan dat je diep hebt moeten zuchten als ik voor de 88e keer dezelfde letters omdraaide. Ik zal de analogieën die je gebruikte om mij een beetje op weg te helpen nooit vergeten. Eerste het bos, dan de bomen. Ik ben altijd onder de indruk van je wetenschappelijke kijk en je talent om de juiste vragen te stellen. Ik hoop dat je dat met nog veel andere studenten mag delen in de toekomst. Het zal niet nieuw (new?) zijn dat je wordt bedankt in een proefschrift. Het is zeker waar (true?) dat ik zonder jou niet zo trots kon zijn op dit eindresultaat. Dus (sowhat?) dankjewel!

Geachte prof. dr. Van Gils en dr. Veldhuis, beste Carla en Wouter. Bedankt dat ik mee mocht werken in de DENSE trial, dat jullie altijd mee wilden denken over de richting van het onderzoek en de klinische relevantie nooit uit het oog verloren.

Ik wil graag alle leden van de beoordelingscommissie, prof. Paul van Diest, prof. Daniel Oberski, prof. Peter Bosman, prof. Chrit Moonen en MD. Moman, van harte bedanken voor uw interesse in mijn werk en de discussie die ik hierover met u mag voeren.

Al mijn hoofdstukken waren niet compleet geweest zonder de inzet van de coauteurs, bedankt voor jullie bijdragen. In het bijzonder wil ik Bas, Marije en Jelmer bedanken voor jullie enthousiasme en goede ideeën.

Mike en Hui-Shan, wat een geluk om mijn kantoor te delen met zulke intelligente, maar bovenal heel leuke mensen. Het was altijd tijd voor thee, en dus een kans om van alles te bespreken. Als het kon, zou ik jullie coauteur maken van mijn Thesis. Superleuk dat jullie mijn paranimfen zijn!

Mike, bedankt voor alles wat je hebt uitgelegd en voor alle gezelligheid en goede gesprekken. Hui-Shan, veel vakken hebben we samen gevolgd. Dankjewel dat ik altijd je “huiswerk” mocht gebruiken om het mijne te corrigeren. Adviezen van jou kan ik altijd klakkeloos overnemen. Jammer dat je me niet eerder hebt geadviseerd mijn computer niet om te gooien.

Dear colleagues of the OIO steeg, thanks so much for all the fun and good talks we had besides our high level research.

Mark, Max, bedankt dat jullie me nog tolereerden op de kamer. Max, heel veel succes op je weg om radioloog te worden. Je trekt altijd je eigen plan en zorgt er vooral ook altijd voor dat je je doel behaalt, dus dit zal ook wel lukken! Mark, trotse Zeeuw, bedankt voor het supereenvoudig toegankelijk maken van een oneindige hoeveelheid computerkracht. Zonder jou was deep-learnen altijd iets te ingewikkeld gebleven.

Gert, Gijs, Simone, Jeroen, Denise, Pascal, Maria, Sieger en Jacobien, bedankt voor jullie interesse en nuance. Ik weet dat mijn PhD te lang een gespreksonderwerp is geweest, maar jullie waren altijd geïnteresseerd. Laten we nog vaak lachen om mijn oneindige PhD-traject tijdens weekendjes fietsen of tijdens het vertellen van sterke verhalen.

Ernst-Jan, als wij samen zouden bepalen hoe onderzoek eruit zou zien, dan weet ik zeker dat onderzoek veel effectiever zou zijn;). Dankjewel dat we altijd konden sparren over de ellende tijdens een PhD-traject en onze successen konden delen.

Aart, Petra, Dirco, Wietske, Mariët en Mark, eindelijk is het zover, ik ben het slimst. Hoewel de één beter begrijpt wat ik aan het doen was dan de ander, waren jullie altijd oprecht geïnteresseerd, dankjulliewel.

Papa en Mama, bedankt dat jullie er altijd voor hebben gezorgd dat ik kon groeien. Altijd zijn jullie geïnteresseerd en staan jullie klaar. Ik begrijp dat het soms moeilijk

was om te horen hoe onzeker het leven van een PhD-student is en dat er ook veel tegenslagen waren. Toch hebben jullie altijd geloofd in een goed einde, bedankt.

Lieve Lianne, zonder jou was het nooit gelukt. Bedankt dat ik altijd alles met je kon vieren of delen. Jij begrijpt altijd precies wat me bezighoudt en wat ik nodig heb. Na een zware week eindigen met pannenkoeken en keukenkaraoke benaderde perfectie.

Lieve Jenthe, na mijn PhD ga ik samen met jou en Lianne verder met de wereld ontdekken.

List of publications

Accepted:

Erik Verburg, Jelmer M. Wolterink, Stephanie N. de Waard, Ivana Išgum, Carla H. van Gils, Wouter B. Veldhuis, Kenneth G.A. Gilhuijs, *Knowledge-based and deep learning-based automated chest wall segmentation in magnetic resonance images of extremely dense breasts*, *Medical Physics*, 2019, Volume 46 (4405-4416)

Erik Verburg, Carla H. van Gils, Bas H.M. van der Velden, Marije F. Bakker, Ruud M. Pijnappel, Wouter B. Veldhuis, Kenneth G.A. Gilhuijs, *Deep Learning for Automated Triaging of 4 581 breast MRIs from the DENSE Trial*, *Radiology*, 2021

Erik Verburg, Carla H. van Gils, Marije F. Bakker, Max A. Viergever, Ruud M. Pijnappel, Wouter B. Veldhuis, Kenneth G.A. Gilhuijs, *Computer-Aided Diagnosis in Multiparametric Magnetic Resonance Imaging Screening of Women With Extremely Dense Breasts to Reduce False-Positive Diagnoses*, *Investigative Radiology*, 2020, Volume 55 (438-444)

Erik Verburg, Carla H. van Gils, Bas H.M. van der Velden, Marije F. Bakker, Ruud M. Pijnappel, Wouter B. Veldhuis, Kenneth G. A. Gilhuijs, *Validation of combined Deep-learning Triaging and Computer-aided Diagnosis in 2,901 Breast MRI Examinations from the Second Screening Round of the DENSE Trial*, *Investigative Radiology*, 2023

Hui Wang, Bas H.M. van der Velden, Max A.A. Ragusi, Wouter B. Veldhuis, Max A. Viergever, **Erik Verburg**, Kenneth G.A. Gilhuijs Kenneth, *Toward Computer-Assisted Triaging of Magnetic Resonance Imaging-Guided Biopsy in Preoperative Breast Cancer Patients*, *Investigative Radiology*, 2021, Volume 56 (442-449)

Laura G. Merckel, **Erik Verburg**, Bas H.M. van der Velden, Claudette E. Loo, Maurice A.A.J. van den Bosch, Kenneth G.A. Gilhuijs, *Eligibility of patients for minimally invasive breast cancer therapy based on MRI analysis of tumor proximity to skin and pectoral muscle*, *The Breast Journal*, 2018, Volume 24 (501-508)

Curriculum Vitae

Erik Verburg (Zwolle, 1 March 1986) started his studies Clinical Technology at university of Twente in 2004, after completing high school at the Heerenlanden College in Leerdam. After obtaining his Bachelor's degree, he became board member in the SportRaad UT. During his Master, Robotics and Imaging, he specialized in image processing. His master thesis on MRI guided HIFU treatment of breast cancer with adjuvant MRI guided radiotherapy was performed at the Image Sciences Institute. After graduation in 2011, he worked for four years at Ophtec BV as a senior development engineer. He was responsible for the development of new optics to be used in implantable intraocular lenses and the implementation of new manufacturing processes. In 2015 he started as a PhD candidate in the group of dr. Kenneth Gilhuijs, working on the DENSE Trial a project funded by the Dutch Cancer Society (KWF, grant number UU-2014-7151), aiming to reduce false-positive follow up and performing breast cancer risk prediction using advanced breast MR image processing. The results of this research project are presented in this dissertation.

