Towards an independent observer of screening mammograms: detection of calcifications

Jan-Jurre Mordang

This book was typeset by the author using LaTeX2$_\varepsilon$.

ISBN: 978-94-92896-79-7

Printed by Ipskamp, Nijmegen

Towards an independent observer of screening mammograms: detection of calcifications

Proefschrift

ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus prof. dr. J.H.J.M. van Krieken,
volgens besluit van het college van decanen
in het openbaar te verdedigen op dinsdag 11 december 2018
om 12:30 uur precies

door

Jan-Jurre Mordang

geboren op 8 april 1988
te 's-Hertogenbosch

Promotoren:	**Prof. dr. N. Karssemeijer**
	Prof. dr. G.J. den Heeten
	AMC Amsterdam
Co-promotor:	**Dr. M. Broeders**
Manuscriptcommissie:	**Prof. dr. E. Marchiori**
	Prof. dr. P.J.F. Lucas
	Dr. M. Lobbes
	Maastricht Universitair Medisch Centrum

The research described in this thesis was carried out at the Diagnostic Image Analysis Group, Radboud University Medical Center (Nijmegen, the Netherlands).
This work was funded by grant KUN 2012-5577 of the Dutch Cancer Society and supported by the Foundation of Population Screening Mid West.

Financial support for publication of this thesis was kindly provided by the Faculty of Science, Radboud University Nijmegen.

Table of Contents

II Evaluation of CAD and breast cancer screening

Introduction

1

1.1 Breast cancer

Cancer is the most deadly type of disease in the world with an estimated 8.2 million cancer-related deaths and 14.1 million new cases worldwide in 2012. An overview of the estimated number of deaths and new cases for the different kinds of cancers is shown in Figure 1.1. Of all cancers, breast cancer is one of the leading causes of death and has the highest incidence in women[1,2]. In the Netherlands, more than 14,500 women are diagnosed with invasive breast cancer and more than 2,000 women with Duct Carcinoma In-Situ (DCIS)[3,4] Furthermore, more than 3,200 women decease because of breast cancer each year[3,4]. By the year 2020, the incidence is estimated to increase to 17,500 new cases, while the breast cancer-related deaths will decrease gradually due to detection of breast cancer at an earlier stage with breast cancer screening and improved treatment of breast cancer[5–8].

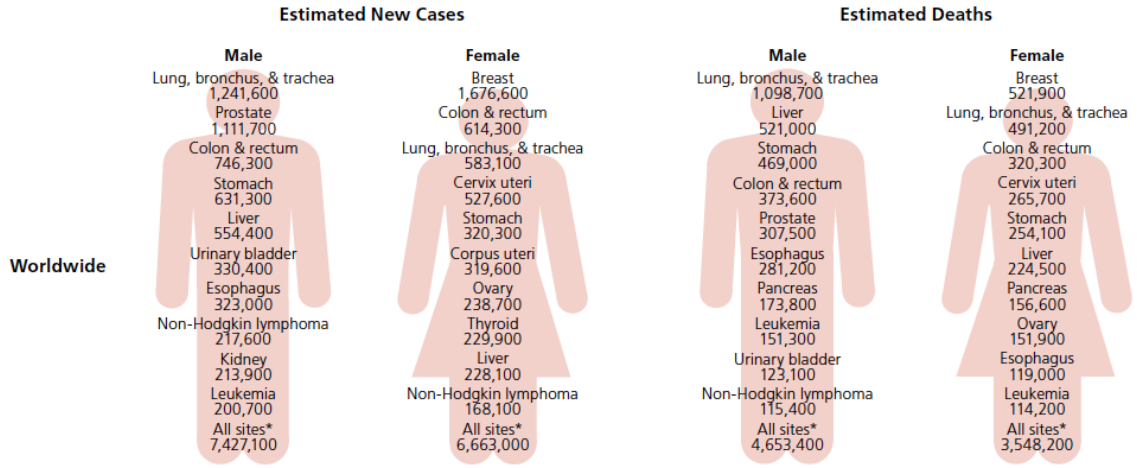


Figure 1.1: Estimates of new cancer cases and deaths in 2012 (**source:** GLOBOCAN 2012[1]).

Malignancy can grow within all types of breast tissue, but in the classical sense, breast cancer originates in either the milk-ducts, in which breast milk is transported to the nipple, or the lobules, where breast milk is produced. A schematic drawing of the anatomy of the breast is shown in Figure 1.2. Several types of breast cancer can arise in other parts of the breast as well, but are less common (<8% of all breast cancers). As long as the cancers cells remain confined within the basal membrane of the ducts, they are defined as DCIS and are not yet harmful although the cancer cells as such can be malignant[9]. When the cancer invades into the stroma (i.e. the surrounding tissue), the cancer becomes invasive and harmful to the patient[10]. After invasion of the stroma, a significant part of the cancers can progress to the lymph nodes and can metastasize further into the body.

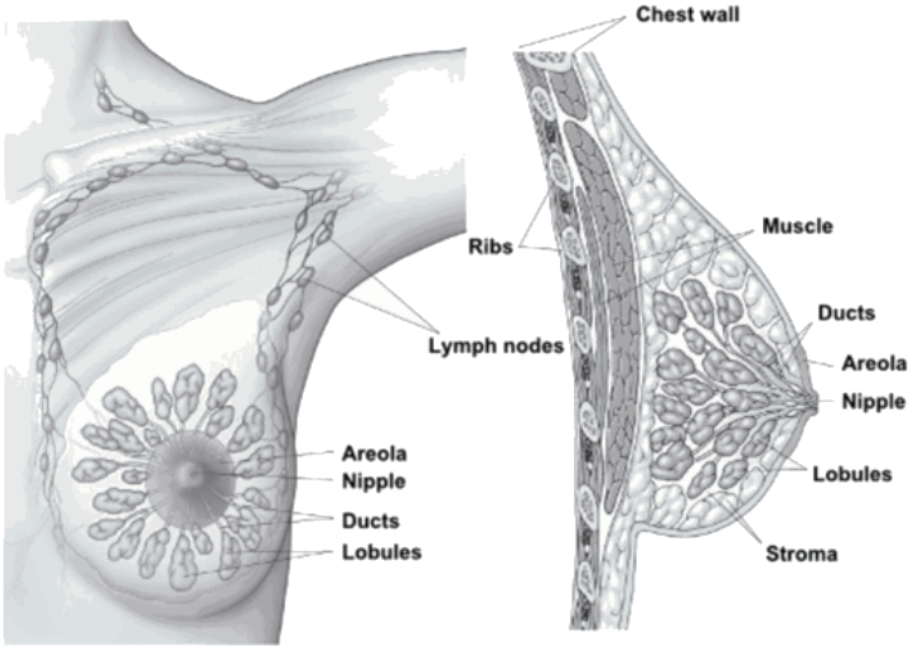


Figure 1.2: Anatomy of the breast (**source:** cancer.org).

An important measure for breast cancer is the grade of the cancer cells in the breast. Grading can be divided into three grades, where the highest grade (grade III) is the most aggressive form of cancer cells[11]. Low-grade (grade I) DCIS cells are slow growing cancer cells and can grow so slowly that they, in general, will not become invasive during the patients lifetime. In contrary to low-grade DCIS, high-grade DCIS cells tend to grow more quickly and, therefore, patients with high-grade DCIS have a higher risk to develop an invasive cancer. Additionally, the recurrence of breast cancer within five years is more likely for high-grade DCIS compared to low-grade DCIS[12]. In general for each individual breast cancer lesion, although the breast cancer stage (specified with the TNM staging method[13]) can change over time, the grading remains the same during the existence of the tumor.

1.1.1 Calcifications

The earliest radiological manifestations of DCIS are calcifications. These small calcium deposits originate in the ducts and lobules and can be a signal of malignancy. However, not all calcifications that can be found in the breast accompany malignancies. Various types of benign calcifications can arise at different locations within the breast and each type can have a different origin and appearance. These types are not related to breast cancer. In Figure 1.3, an overview is shown of typically benign types of calcifications. Although these calcifications are benign in terms of breast cancer, the presence of benign calcifications can be a sign for other types of patholo-

gies. For instance, various studies have analyzed the relation between breast arterial calcifications (BACs), i.e. calcifications located in the blood vessel walls, and cardiovascular disease.[14–25]. Furthermore, BACs have also been related to several other pathologies[26–30].

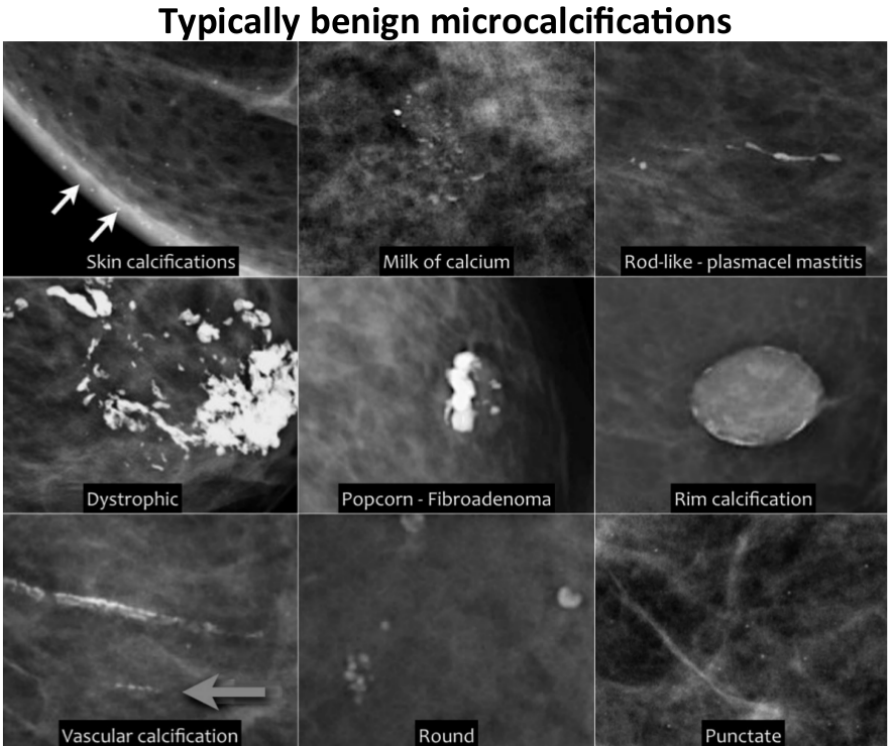


Figure 1.3: Typically benign calcifications that can be present in a mammogram (**source:** radiologyassistant.nl).

For malignancy, suspicious calcifications can be identified by their individual morphology and their distribution within the breast. In Figure 1.4, an overview is given of possible types of suspicious calcifications. Calcifications are suspicious when they have an amorphous or coarse heterogeneous shape. However, although these calcifications are suspicious, they are not always malignant. Calcifications which are fine pleomorphic, thin, linear or curvilinear irregular calcifications (also known as fine-linear, or fine-linear branching calcifications), are considered to have a high probability of malignancy[31].

1.1.2 Soft-tissue lesions

When DCIS develops into an invasive cancer, the breast cancer becomes a soft-tissue lesion, which is the term for masses, architectural distortions and asymmetrical densities within the breast. Most soft-tissue lesions have the main appearance of masses and consist of cancer cells that are more densely packed together and invades the

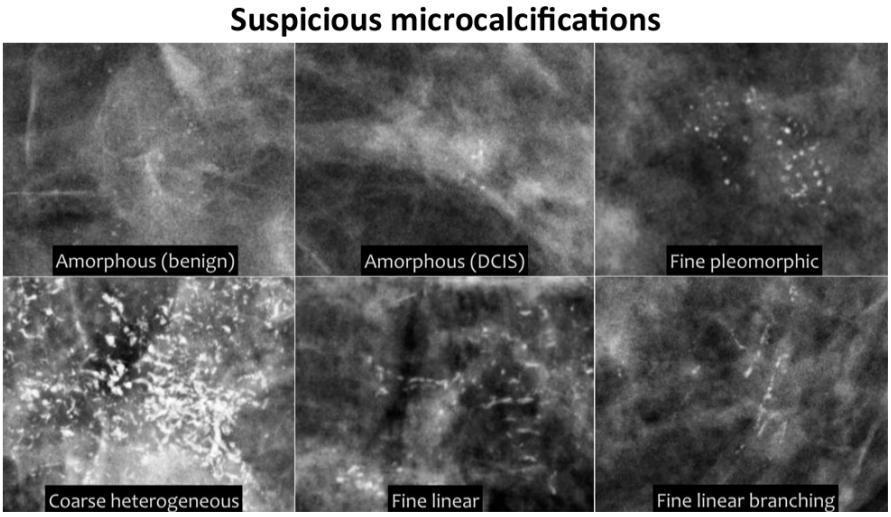


Figure 1.4: Suspicious calcifications that can be present in a mammogram (**source:** radiologyassistant.nl).

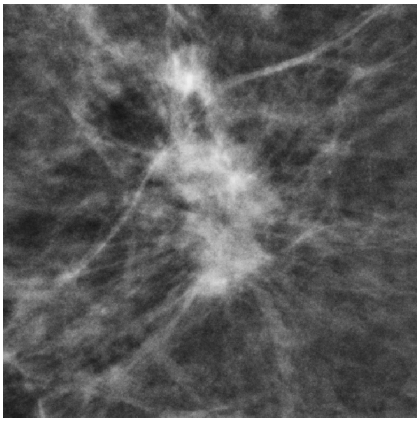


Figure 1.5: Example of a malignant soft-tissue lesions in mammography.

surrounding tissue, which consists mainly of fatcells and fibrous tissue. The boundary of this type of lesion can vary between circumscribed, indistinct or spiculated. The latter type, are stellated patterns of lines that are directed towards the center of the mass. These spiculations are an important sign for malignancy of the lesion. Architectural distortions are a disruption of the normal pattern in the breast without a visible mass and are less often an invasive cancer. The asymmetrical densities, a mismatch between the density pattern between the left and right breast or acquisitions at different view angels of the breast, are also less often a malignancy. An example of a malignant soft-tissue lesion with spiculations is shown in Figure 1.5.

1.2 Breast cancer screening

Breast cancer is best treatable when detected as early as possible[32,33]. Therefore, early detection of this disease is essential to decrease breast cancer mortality[34,35]. To accomplish this, breast cancer screening programs are implemented in many developed countries where women from a certain age are periodically invited for a breast cancer examination. It is important to underline that the positive effects of screening are mainly due to the principle of repetition. For this reason, the first round should be considered differently from the repeated rounds and monitored and reported separately. Between 20% and 50% of non-palpable DCIS develops into invasive breast cancer[36–38], and because non-palpable DCIS can only be detected with medical imaging, screening examinations are done with mammography. This approach has proven to be a cost effective measure for the early detection of breast cancer[39,40].

In the Netherlands, women between the age of 50 and 75 are biennially invited for a breast cancer screening exam. During this exam, a full-field digital mammogram (FFDM) is acquired with a mammography system. An example of a mammography system is shown in Figure 1.6. Each mammogram consists of two mammographic images of the left and right breast that are acquired during one screening examination. The two views that are acquired are called: the Medio-Lateral Oblique (MLO) and the Cranio-Caudal (CC) view. The MLO is acquired with the X-ray tube rotated 45 °medially and the CC is acquired from the top of the breast along the line of gravity, respectively[41]. A schematic drawing of both views is shown in Figure 1.7. The four acquired images are read by two in principle independent readers. In this double reading, two radiologists assess the mammogram and score both breasts consecutively. Scoring of the mammogram is performed according to the Breast Imaging-Reporting And Data System (BI-RADS)[31,42] in which guidelines are presented to consistently score and report breast cancer lesions[41].

1.3 Computer-aided detection

The advancement of medical imaging over the past decades results in a tremendous amount of medical images, substantially increasing the workload of radiologists[43–47]. Especially in screening programs such as breast cancer screening, where millions of medical images are acquired each year. Besides the increasing workload, the interpretation of medical images is subjective to the individual skills of (screening) radiologist and depends also on experience and their compliance with reporting guidelines. For consistent reporting different reporting systems are available for spe-

Figure 1.6: An example of a mammography system, commonly used in breast cancer screening.

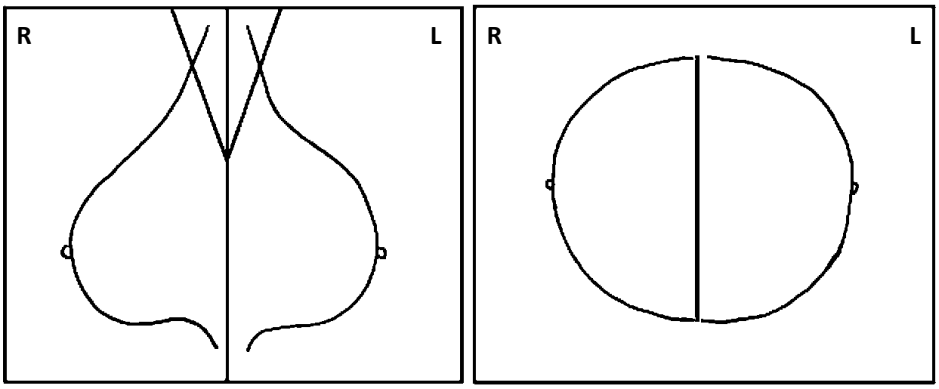


Figure 1.7: A schematic drawing of the the Medio-Lateral Oblique and the Cranio-Caudal views acquired during a breast cancer screening exam.

cific diseases and modalities such as the Lung-RADs[48], the PI-RADS[49], and the BI-RADS[31] for lung, prostate, and breast imaging, respectively. The difference in reading quality between radiologists can result in a difference in the diagnosis of a patient and can have a big impact on the number detected cancers. For instance, a sensitivity difference varying between 18% and 40% has been observed when mammograms are read by individual breast cancer screening radiologists[50–53]. Therefore, in many European countries, double reading has been introduced to reduce the variability in breast cancer screening performance and to increase sensitivity[53–56]. However, double reading demands additional radiologists which increases their workload even more and increases costs.

To reduce the radiologists' workload and to improve quality of reading medical images, Computer-aided detection (CAD) systems have been developed and have been extensively explored for the past decades[57–59]. In these systems, various (semi-)automatic algorithms are used to analyze medical images and give a response to aid the radiologist. In general, there are two types of responses and, consequently, two types of CAD systems. In a CADe (computer-aided detection) system, the general aim of the system is to detect abnormalities in medical images. Therefore, the output of a CADe system are marks (or findings) of potential locations of abnormalities within the image. This type of system is mainly used to reduce the number of abnormal regions that could potentially be overlooked by the radiologist. Additionally, many of these systems supply a score with the supplied findings to show how certain the system is about a specific location to be abnormal. The second type of CAD systems are Computer-aided diagnosis (CADx) systems. These systems are developed to be an aid for the radiologist in the interpretation of abnormal regions. For example, CADx systems can help in the interpretation and classification of benign and malignant disease in various diseases and imaging modalities such as lung cancer in CT-scans[60], breast cancer in mammography[61,62], and prostate cancer in MRI and ultrasound imaging[63].

The implementation of a CAD system into the daily workflow of radiologists can be done with different setups. For instance, a CAD system can be used during reading of medical images where usage of the system is regulated by the radiologists themselves. In this setup, the aid of the system can be either aimed at the detection of abnormalities (CADe) or as a interactive decision support for the evaluation of found abnormalities (CADx). In a detection support system, CAD findings can be prompted on the image when desired to check if certain regions were not overlooked. In the decision support system, CAD findings are only shown when the radiologists wants to know if a certain region is found to be suspicious by the CADx system[64,65]. In another setup, a CAD system can be used as a completely in-

dependent reader of medical images (also mentioned as CADr[66]). For instance in mammography screening, the system can be used as a first reader where only images with abnormalities are shown to the radiologist or as a second reader were the images are checked for potentially overlooked abnormal regions. Furthermore, the system can be a substitute of one radiologists in double reading.

1.3.1 Mammography CAD in breast cancer screening

The implementation of a CAD system in mammography screening is not a novel idea and, in the United States, CAD systems are already widely used in screening practice for almost two decades[67]. Since then, a lot of effort has been put in further development of these systems to approach human performance in reading mammograms. A mammography CAD system generally consists of two separate CAD systems: a system for detecting calcifications and a system for detecting masses. As this thesis mainly focusses on calcification CAD systems in breast cancer screening, we will discuss this type of system more thoroughly and the mass CAD system will only be discussed briefly.

Calcification CAD system

In the past decades, much research has been done in the development of calcification CAD systems and it is still a prominent research subject to this day[61,62,68–72]. The main reason why this development is still ongoing is due to the number of false positive locations that are marked by the system which is still around 100 times higher than the number of false positives marked by radiologists in screening[73]. The baseline CAD system used in this thesis is based on a calcification CAD system developed by Bria et al[74] (2013). In their paper, a calcification CAD system was presented that can compete with high-end commercial CAD systems. The strategy for the detection of suspicious calcifications in mammograms consists of four main steps: 1) the image will be processed with a filter to increase and equalize the contrast in the image, 2) a pixel detector is applied to detect pixel-candidates for potential calcifications, 3) calcifications are segmented and clustered based on the mapping of potential candidates and, 4) a cluster classifier is applied on the clustered calcifications to reduce the number of false positive clusters. These steps of the baseline CAD system will be discussed in the next subsections.

1) Pre-processing - The first step in the whole CAD pipeline is the segmentation of the breast such that no other structures are present in the image[75]. Furthermore in digital mammography, a dominant source of noise in the mammograms is quantum noise which is unavoidable in X-ray imaging[76]. However, a filter can be applied to

equalize the noise across the image. Together with a rescaling of the pixel values, the following formula can be used:

$$T(y) = \frac{T_{max}}{\sqrt{y_{max}}}\sqrt{y} \tag{1.1}$$

where, T_{max} is the maximum intensity of the transformed scale, e.g. for a 16 bits/pixel image this would be 65,535, y_{max} is the maximum intensity of the original mammogram, e.g. 14 bits/pixel (y_{max}= 16,363) and y is the value of each individual pixel in the image.

2) Calcification candidate detector - The main challenge for a pixel classifier in detecting locations of calcifications is the huge class-imbalance between pixels belonging to calcifications and pixels belonging to other breast tissue, considering that calcifications can consist of only few pixels (∼10 pixels) where a group of only three calcifications can already be suspicious and the millions of pixels belonging to other breast tissue. Additionally, a pixel classifier should be able to process millions of pixels in a reasonable time as the system, when implemented in breast cancer screening, should be able to process millions of mammograms yearly. Consequently, the prerequisites for a suitable pixel classifier are that it should be able to overcome a large class-imbalance and cannot be too complex.

A suitable classifier is the cascade classifier[77] which consists of series of nodes containing very simple classifiers that are subsequently applied on fast computable features. The rationale behind the cascade is that pixels that would be easily dismissed as calcifications will be removed from the candidate list by one of the first nodes. Classifiers in the subsequent nodes can focus on the differentiation between calcifications and more difficult samples of other breast tissue. The classifier in our cascade classifier is a GentleBoost classifier[78] which is trained on Haar-like features[79,80]. To classify each pixel in the image, a nxn pixel patch (e.g. 13x13) is extracted for all pixels. On each patch, various Haar-like features groups are calculated at various scales and locations within the patch. Examples of the Haar-like feature groups are shown in Figure 1.8. In each node of the cascade, pixels are classified and pixels with a classification score below a certain threshold are removed for further classification in the subsequent nodes. In the last node, each remaining pixel, i.e. a pixel classified as a positive pixel in all nodes, receives a final classification score. This score is the output of the last classifier, all other pixels receive the value zero. An example of the output of the classifier is shown in Figure 1.9(b).

Several parameters should be set while training the GentleBoost classifiers in each node. The most important are the minimum detection rate and maximum false positive rate of the individual classifiers in the cascade. Sensible values for these two settings would be a detection rate of 0.99 and a false positive rate of 0.3 which for

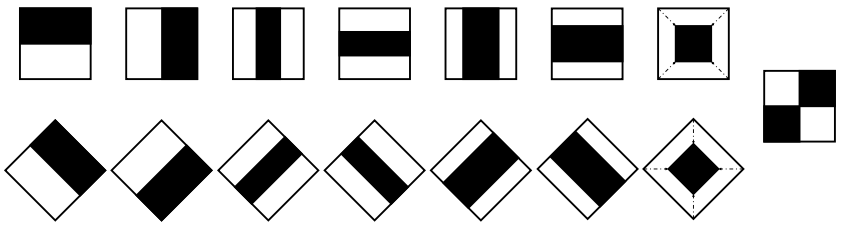


Figure 1.8: Haar-like feature groups, each feature group consists of features at all possible scales and translations within an nxn pixel patch.

a cascade with 5 nodes would yield (in theory) a detection rate of $0.99^5 = 0.95$ and a false positive rate of $0.3^5 = 0.002$. During training, Haar-like features are selected by the GentleBoost classifier until these two criteria are met. Because in the beginning very easy differentiable samples can be filtered out, the first nodes will need fewer selected features compared to the later ones. The final number of nodes can be controlled by several factors such as a pre-specified overall detection rate of the system and false positive rate, the number of training samples that are still available in the later nodes, or when adding nodes does not improve the overall performance of the cascade anymore. Training of the cascade classifier takes more time than applying it to new images because, during training, all Haar-like features have to be calculated for all samples. For example, around 40,000 samples are used (with a class-imbalance of around 1:5) for training each node and for a single patch of 13x13 more than 45,000 Haar-like features are calculated. While when the cascade classifier is applied to a new image, only few Haar-like features are calculated ($<$10 features in the first node up to $<$50 in the later nodes) which are quite fast to calculate ($\sim$10 seconds for a whole image).

3) Calcification segmentation and clustering - After the candidate detector, calcifications are segmented and clustered together in groups. To segment calcifications, a connected component analysis is applied on the output image of the candidate detector resulting in a set of components. Then, a threshold is set (T_{calc}) and components with less than 2 pixels above T_{calc} are removed and the remaining components are defined as detected calcifications. To remove macrocalcifications, which are typically benign, bounding boxes are calculated for each detected calcification and calcifications with a bounding box that is larger than 1mm in both horizontal and vertical direction are removed. Additionally, detected calcifications that have an overlapping bounding box with or lie within 2 pixels of a macrocalcification are removed as well. The remaining detected calcifications are then clustered together

by grouping them according to their distance to one another where the maximum distance between individual detected calcifications is set to 10mm. An example of groups with detected calcifications is shown in Figure 1.9(c).

1

4) False positive removal - Not all groups with detected calcifications will contain suspicious calcifications but also typically benign calcifications or noise might be detected. Therefore, to reduce the number of false positives, a cluster classifier is trained on detected groups of calcifications to differentiate suspicious calcifications from benign calcifications. This classifier uses various features, which are calculated either at an individual calcification level within the groups and at the level of the whole groups itself. These features are based on shape, topology, probability, and texture and are described in Bria et al[74] (2013) and Veldkamp et al[81] (1999). After classification, each detected calcification group receives a score reflecting how confident the system is that the group is suspicious. In Figure 1.9(d), an example is shown of the final output of the baseline calcification CAD system.

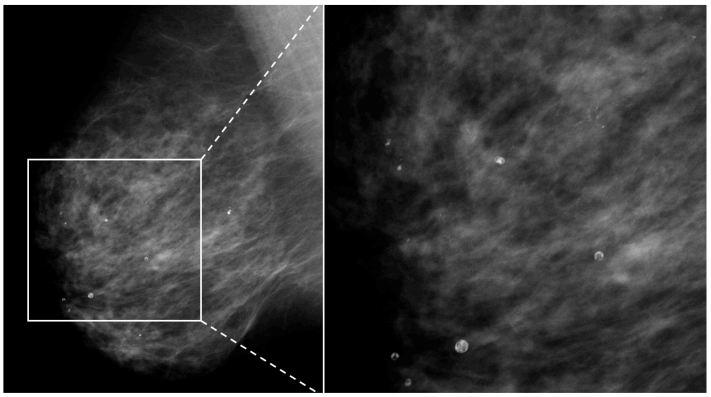
Soft-tissue lesion CAD system

The development of CAD systems for the detection of soft-lesions is still an ongoing research field[82–88]. The soft-tissue lesion detection system used in this thesis has the following design. After pre-processing the image, each pixel in the image is analyzed by a pixel classifier to generate a likelihood image for potential candidates[83]. To obtain the location of potential soft-tissue lesions, local maxima are determined in the likelihood image and the lesions are segmented with dynamic programming[89]. Additionally, patches are made from the segmented lesions. These patches, together with a list of features calculated on the segmented lesions[83], are classified with a convolution neural network (CNN)[88]. The output of the soft-lesion CAD system is a segmented lesion together with a suspiciousness score supplied by the CNN. Examples of the intermediate results are shown for respectively the output of the candidate detector and the lesion segmentation.

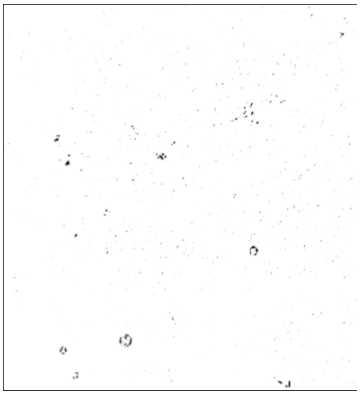
1.3.2 Evaluation of CAD

When a new CAD system is developed, its performance should be validated. The validation strategy should be done as properly as possible to be able to compare the performance of the new proposed system to other CAD systems. In this section, the evaluation methods are described which are used throughout this thesis.

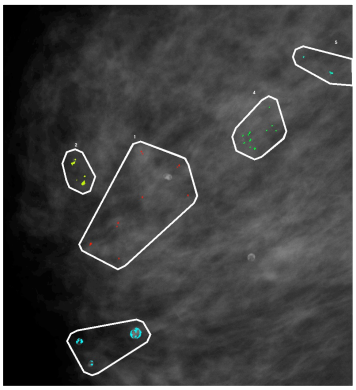
Commonly, a CAD system is validated on a reference dataset (or ground truth) that is created based on the diagnostic findings of radiologists and the histopathological findings after diagnostic follow up. Based on these findings, annotations are



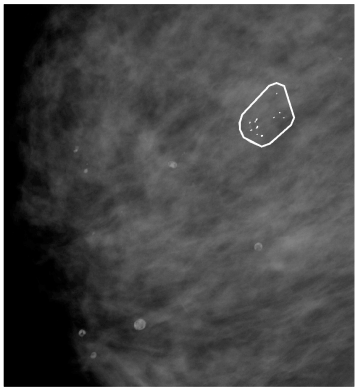
(a)



(b)



(c)



(d)

Figure 1.9: Examples of intermediate results of the calcification CAD system; (a) the input screening mammographic image and a zoomed in region, (b) the calcification candidate selection, the black pixels represent locations of the calcification candidates, (c) the detected groups after calcification segmentation and clustering, and (d) the remaining calcification group after the false positive removal.

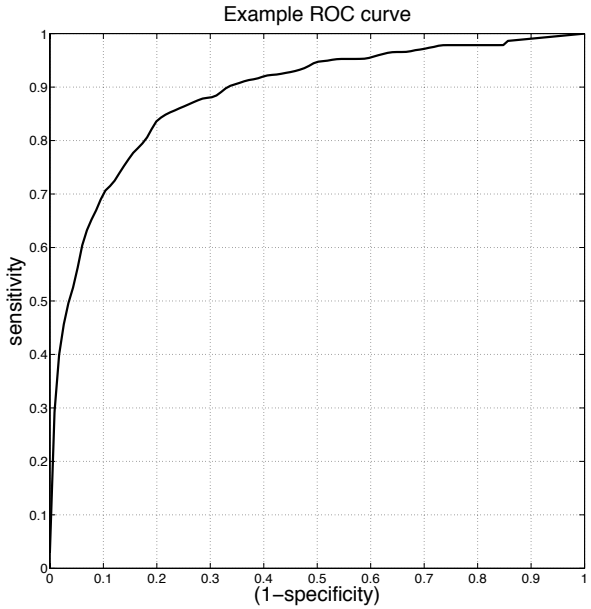


Figure 1.10: An example of a ROC curve.

drawn capturing the malignant lesion (i.e. calcifications or soft-tissue lesions) in the image. The CAD system is applied on this dataset and the percentage of detected malignancies (the true positive rate or sensitivity) and the percentage of detected normals (the false positive rate or 1 - specificity) are calculated. Because the findings produced by the CAD system have a classification score, Receiver Operating Characteristics (ROC) analysis can be performed. With ROC analysis, all samples in the dataset are ranked according to their classification score. To obtain an ROC curve, various thresholds (T_{group}) are set on the classification scores and at each threshold the number of true positives (detected malignancies) and false positives (detected non-malignancies) are calculated to determine the operating point, i.e. the combination of the sensitivity and specificity at a given T_{group}. When various thresholds are set, various operating points can be calculated and a ROC curve can be plotted, an example of an ROC curve is shown in Figure 1.10. Often the Area Under the Curve (AUC) of the ROC curve is calculated to give an overall metric of the performance. The value of the AUC lies in the range of 0 and 1 where an AUC of 1 means perfect performance, i.e all malignancies are detected without any false positives.

However, to obtain a ROC curve it should be specified clearly when a finding of the CAD system is a true positive or a false positive, i.e. a detected malignancy and

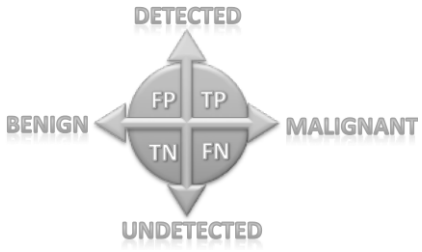


Figure 1.11: The four measures to calculate specificity and sensitivity.

a detected non-malignancy, respectively. In Figure 1.11, a schematic visualization is shown for the definitions of the CAD findings. Various criteria can be specified for the definition of a true positive and depend on the evaluation of the CAD system. In this thesis, two different criteria were used: 1) a finding is defined as true positive when at least two detected calcifications are located within the reference annotation and 2) the distance between the center of the CAD finding and the center of the reference annotation are used where a finding was marked as a true positive when this distance was <10mm. Throughout this thesis, a false positive is defined as each finding that is detected in the images of an exam that has not been recalled in screening (a normal exam), assuming that these images do not contain any abnormalities. Moreover, often exam-based ROC analysis is performed. When exam-based ROC analysis is carried out, the same definitions for true positive and false positive findings are used. However to obtain exam-based scores, the true positive with the highest score in all images of the exam was taken for exams with a malignancy. For the normal exams, the finding with the highest score in all images of an exam was taken as the false positive score.

Furthermore, another analysis that can give a good insight in the performance of a CAD system is Free-response ROC (FROC). Similar to ROC analysis, the number of true positives and false positives are calculated at various classification scores and the definitions are the same. However, instead of plotting the sensitivity in terms of the specificity, it is plotted in terms of the number of false positives per (normal) image (FP/I). In this thesis, the sensitivity for FROC analysis is calculated exam-based, identical as in ROC analysis, and the number of false positives per image are calculated by taking all findings detected in normal images and ranking them to their classification score. To obtain the FROC curve, T_{group} is set at various values and for each value the number of false positives is determined and divided by the total number of normal images in the test set. An example of an FROC curve is shown in Figure 1.12, in this graph it can be seen that for instance at a sensitivity of 88%, the CAD system produces one false positive in every five images (FP/I = 0.2). Note that FROC curves are commonly plotted on a logarithmic scale. Calculating these

operating points for a FROC curve makes it possible to see how the CAD system would fit in a screening environment as a CADe system because it directly shows the number of false positive marks it will generate at a certain sensitivity.

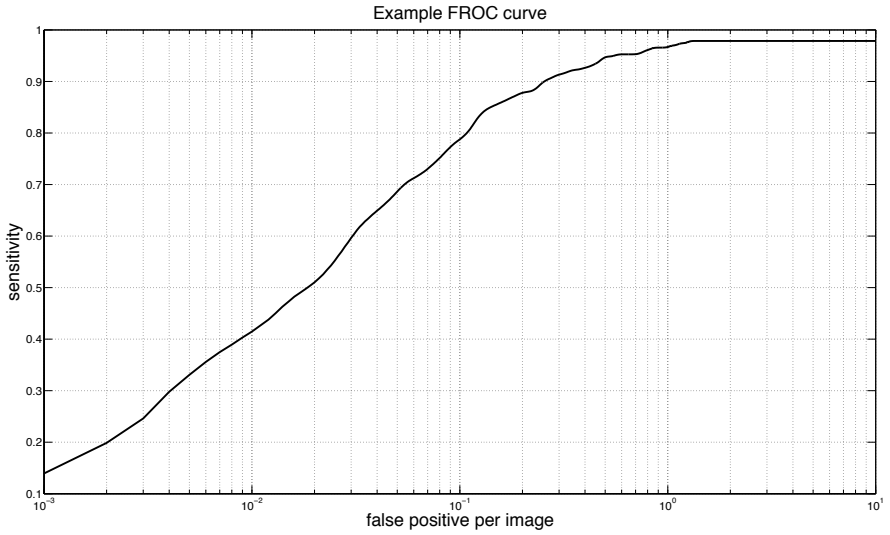


Figure 1.12: An example of a FROC curve.

Besides directly comparing (F)ROC curves between different systems, a statistical comparison is also relevant for evaluation. To compare two systems bootstrapping can be used[90–93]. Bootstrapping is a non-parametric method to obtain confidence intervals for each (F)ROC curve. The bootstrapping method consists of resampling the reference dataset with replacement n-times (commonly, n>1000), for each sampled set an (F)ROC curve is calculated for each system and are compared. Statistical comparison is often done by calculating the AUC of the (F)ROC curve for each sampled set and derive the p-value from the differences in AUC. Moreover, when a certain range is of interested, e.g. the high specificity range between 0.8 and 1.0, the partial AUC (pAUC) can be calculated for each bootstrap and compared between systems. In general, it is assumed that two systems are statistical different at a $p<0.05$.

1.4 Thesis outline

Several reader studies have shown that the detection rate of individual radiologists increases when a CADe system is used[73,94–96], at the cost of a slight increase in recall rate[73,97–100]. However, there is no convincing evidence yet that the incorporation of CAD systems into the mammography reading workflow contributed to an overall improvement of screening performance in daily practice[98,101]. It is not clear why there is a discrepancy between the positive results of reader studies where a CAD system was used in reading sessions compared to the less positive results observed

by incorporating a CAD system in the daily work flow of breast cancer screening.

An important factor that can play a role in this disappointing result might be explained by the relatively high number of false positives that are marked by CAD systems. These false positives arise because a high sensitivity is desired to ensure that lesions are not missed and, consequently, CAD systems have to operate at a low specificity, which is not yet optimal for current systems. The high number of false positives can lead to an increase in the number of women being unnecessarily recalled for a clinical follow-up[73,96,97], an increase in interpretation time of the mammograms[102], and a loss in confidence in the CADe system[102]. All of these drawbacks can have a negative effect on the usage of a CAD system and can result in a refusal of using it during the daily work flow of screening. Consequently, it is difficult to assess if CAD marks are actually judged by the radiologist or generally ignored when CAD is available to them and the usage CAD is reimbursed[103]. Therefore, instead of using a CAD system as an aid for radiologist during mammography screening, a better solution might be to use the system as a completely independent observer. In this setup, the CAD system can serve as a stand-alone system and its output can be combined with the grading of the radiologists. However, before employing CAD as an independent observer, the number of false positives should still be reduced and should be compared to the performance of radiologists.

Therefore, the main goal in this thesis is to reduce the number of false positives produced by the calcification CAD system to achieve a comparable performance as achieved by screening radiologists. This thesis consists roughly of two parts. In the first part, improvements on the baseline CAD system will be discussed and in the second part, the screening sensitivity of detecting calcifications is assessed and the stand-alone CAD system is evaluated. The first part consists of chapters 2-5: in **Chapter 2**, an improvement is proposed of the initial candidate detector of the baseline system. In this chapter, a deep-learning approach has been implemented where a convolutional neural network was trained and compared to the cascade classifier. In **Chapter 3**, a case-based selection stage has been to select cases with breast-arterial calcifications, one of the most common false positives produces by the CAD system. In **Chapter 4**, a breast-arterial reduction stage has been applied on the selected cases to reduce the number of false positives and increase the performance of the CAD system. In **Chapter 5**, a framework is presented to remove obvious false positives, these are false positive CAD findings that would be easily dismissed by the radiologist. The second part consists of chapters 6-7: in **Chapter 6**, the performance of radiologists in detecting malignant calcifications in breast cancer screening has been assessed where the exams prior to screen-detected and interval cancers have been retrospectively evaluated together with a CAD system. In **Chapter 7**, the calcifica-

tion CAD has been merged with a mass CAD to assess the overall performance of the mammography CAD system as an independent observer. The performance of the stand-alone CAD system is compared to 109 screening radiologists.

Part I

Improvements of the calcification CAD system

A deep learning approach for the detection of calcification candidates

2

Jan-Jurre Mordang, Tim Janssen, Alessandro Bria, Thijs Kooi, Albert Gubern-Mérida and Nico Karssemeijer

Original title: Automatic microcalcification detection in multi-vendor mammography using convolutional neural networks

Published in: Breast Imaging, 2016

Abstract

Convolutional neural networks (CNNs) have shown to be powerful for classification of image data and are increasingly used in medical image analysis. Therefore, CNNs might be very suitable to detect calcifications in mammograms. In this study, we have configured a deep learning approach to fulfill this task. To overcome the large class imbalance between pixels belonging to calcifications and other breast tissue, we applied a hard negative mining strategy where two CNNs are used. The deep learning approach was compared to a current state-of-the-art method for the detection of calcifications: the cascade classifier. Both methods were trained on a large training set including 11,711 positive and 27 million negative samples. For testing, an independent test set was configured containing 5,298 positive and 18 million negative samples. The mammograms included in this study were acquired on mammography systems from three manufactures: Hologic, GE, and Siemens. Receiver operating characteristics analysis was carried out. Over the whole specificity range, the CNN approach yielded a higher sensitivity compared to the cascade classifier. Significantly higher mean sensitivities were obtained with the CNN on the mammograms of each individual manufacturer compared to the cascade classifier in the specificity range of 0 to 0.1. To our knowledge, this was the first study to use a deep learning strategy for the detection of calcifications in mammograms.

2.1 Introduction

Convolutional Neural Networks (CNNs)[104,105] have been shown to be very powerful in the classification of large image databases[106,107]. Moreover, CNNs are also increasingly applied in medical image analysis, e.g. included in CADe systems, and have shown to achieve cutting edge performances[108–110]. Another great advantage of applying CNNs is that the networks themselves determine the most descriptive features to separate the positive from the negative class while current CADe systems use pre-determined features which can lead to a loss of information or an increase of processing time. Therefore, CNNs might be very suitable for detecting calcifications in mammograms.

One of the main difficulties in the detection of calcifications in mammograms is that the positive class, i.e. pixels belonging to calcifications, is very small compared to the negative class, i.e. other breast tissue. This large class imbalance is a big impediment for most classification strategies and make them unsuitable for this task. A previous study about applying CNNs to a class imbalanced dataset yielded very good results[108]. Additionally, a single mammogram consists of millions of pixels to be analyzed. Therefore, a CADe system should operate very fast. A CADe system with high-complexity classifiers and/or features can lead to a very slow system and consequently become useless to process millions of images. An additional benefit of CNNs is that they can be applied with a very high computation speed.

The purpose of this study is to implement and study the performance of a deep learning approach with CNNs to detect calcifications in mammograms. The proposed system is compared to a current state-of-the-art calcification detection approach. Another important aspect for a CADe system to be applicable in breast cancer screening is their compatibility with mammograms acquired with mammographic units developed by different vendors. Each vendor has its own detector type for making mammograms which can result in a substantial variation in noise characteristics and appearance. Therefore, we used a heterogeneous dataset consisting of mammograms acquired with mammographic units developed by three different manufacturers.

2.2 Methods

2.2.1 Materials

For this study, we collected a multi-vendor and multi-center dataset consisting of 490 mammograms acquired with Hologic digital mammography systems (Hologic, Bed-

Dataset configuration			
Manufacturer	# cases	# exams	# mammograms
Hologic	104	132	490
GE	255	268	1044
Siemens	23	23	72

Table 2.1: Overview of the dataset.

ford, Massachusetts, United States), 1044 mammograms acquired with GE Senographe systems (GE, Fairfield, Connecticut, United States), and 72 mammograms acquired with Siemens Mammomat Inspiration systems (Siemens, Erlangen, Germany). In the dataset all available medio-lateral oblique and cranial caudal views of the left and right breast were included. All mammograms were acquired with standard clinical settings and unprocessed raw FFDM images were used in this study. The data acquired with the Hologic digital mammography systems were obtained from women whom participated in a national screening program (Bevolkings Onderzoek Midden-West, The Netherlands) and were referred for diagnostic follow up. The other mammograms were acquired in our own institution after referral in screening. An overview of the dataset is shown in Table 2.1. In all mammograms, individual calcifications were annotated based on the diagnostic reports. Annotations were made by marking the center of each calcification.

2.2.2 Convolutional neural networks

The general aim for a calcification detector is to classify each pixel in the mammogram in one of two classes: calcification or non-calcification. We propose a hard negative mining strategy to overcome the large class imbalance between the calcification and non-calcification classes. First, a CNN is trained on a small dataset. Second, the trained CNN is applied to the whole dataset to remove the easy samples. Finally, a second CNN is trained on a larger dataset which contains the hard negative samples, i.e. samples that are more difficult to differentiate from calcification pixels than easy classifiable samples. For new samples such as those in the test set, only the second CNN is applied.

To train a CNN, patches, sub-images centered around the pixel of interest, are obtained from both the positive and negative class. These patches are then fed into the first convolutional layer. In each layer, filters are trained to divide the data in separate classes. This approach uses all information within the patch that is supplied

to the CNN and determines its most descriptive features between the two groups by itself. Training of the CNN is an iterative process, where in each iteration (or epoch) network parameters and discriminative parameters are optimized. In each epoch, the CNN minimizes a cost function by updating its parameters via back-propagation.

A CNN consist of several layers, the most commonly used layers are the convolutional layers, the pooling layers and the fully connected layers. The convolutional layers consist of a set of learnable 2D rectangular filters. These filters are convolved with the input patch and the activations are passed through an activation function. The pooling layers reduce the spatial size of the input by sub-sampling the output of the previous layer. Commonly, the maximum activation is taken over a sub window with a specified stride and are called max-pooling layers. The third type of mainly used layers are the fully connected layers. These layers have full connections to all activations in the previous layer. This type of layers are commonly used in regular neural networks. The final layer is a fully connected layer with two output neurons, one for each class.

The CNN structure in our study is inspired by the OxfordNet[111]. This structure consists of repetitions of two convolutional layers, with 32 filters each, followed by a max-pooling layer of size 2x2 and a stride of 2. Additionally, fully connected layers are used as final layers and a soft-max function calculates the final output. The two CNNs used in this study consist of 2 repetitions followed by three fully connected layers. An overview of the CNN architecture used for both CNNs used in this study is shown in Table 2.2. To reduce over fitting, dropout is applied for each fully connected layer during training of the CNNs[112].

2.2.3 Cascade classifier

A current state-of-the-art method for the detection of calcifications is the cascade classifier[74]. This cascade classifier consist of a sequence of nodes where in each node an independent, single classifier classifies the patches. In each node, patches with a classification score below a specific threshold, which is determined during training, are filtered out and receive a final score of zero. Patches which remain in the last node receive the score of the last classifier. GentleBoost classifiers are used as single classifiers and regression stumps are used as weak classifiers[78]. These GentleBoost classifiers are trained on straight[77] and 45°rotated[80] Haar-like features.

During training, each GentleBoost classifier is optimized on a validation set. This optimization is based on two criteria: the detection and false positive rate on the validation set. The validation set is classifier by the GentleBoost classier after adding

CNN Architecture				
Layer	Layer Type	Filter Size	Input size	Output size
1	Convolutional	3x3	1x13x13	32x13x13
2	Convolutional	3x3	32x13x13	32x13x13
3	Max-pooling	2x2 (stride = 2)	32x13x13	32x6x6
4	Convolutional	3x3	32x6x6	32x6x6
5	Convolutional	3x3	32x6x6	32x6x6
6	Max-pooling	2x2 (stride = 2)	32x6x6	32x3x3
7	Fully connected	256	32x3x3	256
8	Fully connected	256	256	256
9	Fully connected	2	256	2

Table 2.2: Overview of the convolutional neural network architecture. The input for the convolutional layers are zero padded to preserve the input size.

a weak classifier. If the criteria are not met, another weak classifier is added to the GentleBoost classifier. Training of the node finishes when the criteria are met or the maximum number of weak classifiers is reached. After each node, all samples in the training set are classified by the newly formed node and samples are removed when they received a classification score below the trained threshold. Training of the whole cascade is stopped when there are too few negative samples left in the training set.

2.2.4 Experiments and evaluation

From the dataset two sets were created, a training set and a test set. For each annotation, a patch was extracted from the mammogram and were considered as positives. Each patch had a size of 13 x 13 pixels with the individual calcification centered in the patch. From the Hologic and GE data 80% of these positives were included in the training set. The remaining 20% of the Hologic and GE positives together with all positives obtained from the Siemens data were included in the test set. Additionally, negative patches were randomly taken from all mammograms (excluding the positive locations). For the training set, up top 35,000 negative patches were extracted from each mammogram in the Hologic data and up to 17,500 from each mammogram of the GE data. For the test set, up to 70,000 negative patches were extracted

Samples per dataset				
Manufacturer	Training set		Testing set	
	# Positives	# Negatives	# Positives	# Negatives
Hologic	5,744	13,321,334	1,868	7,244,003
GE	5,967	14,625,465	1,652	6,842,428
Siemens	-	-	1,778	4,234,545
Total	11,711	27,946,799	5,298	18,320,976

Table 2.3: Overview of the training and test sets.

per mammogram in the Hologic and Siemens data and 35,000 per mammogram in the GE data. An overview of all samples in each training and test set is shown in Table 2.3.

The CNNs were trained on the training set. During training, the learning rate was initially set to 0.01 and linearly decreased to 0.0001 over the maximum number of epochs. For the first CNN, the maximum number of epochs was set to 1500 and to 750 for the second CNN. However, an early stopping criterion was set to prevent the CNN from over fitting: when the validation loss did not change over 100 epochs, training of the first CNN was stopped. For the second CNN early stopping was set to 50 epochs.

Training of the first CNN was performed on a balanced dataset containing 10 times the number of positives (each positive was taken 10 times) and an equal amount of negatives randomly sampled from the training set. In total, 117,110 negatives and 117,110 positives were used to train the first CNN. The second CNN was trained on a dataset containing 1 million samples. All 11,711 positive samples were included in this set together with 988,289 negative samples. The negative samples were obtained by weighted sampling of the whole training set according to the classification scores obtained with the first CNN. Furthermore, to create more positive samples for training the second CNN, positive samples were augmented. Augmentation was performed by flipping the positive samples horizontally and vertically and by rotating the patches 90°, 180°, and 270°. In each mini-batch the number of positives and negatives were balanced.

For evaluation, the performance of the (second) CNN was compared to a current state-of-the-art method for calcification detection: the cascade classifier[74]. The cascade classifier consists of several nodes with one single GentleBoost classifier. For training of the cascade classifier, all positives and all negatives in the training set

were used. Subsequently, both systems were applied to the test set and Receiver Operating Characteristics (ROC) analysis was performed to compare the two systems. Furthermore, the mean sensitivity of the ROC curve in the specificity range on a logarithmic scale was calculated and compared. The mean sensitivity is defined as:

$$\overline{S}(i,j) = \frac{1}{ln(j) - ln(i)} \int_i^j \frac{s(f)}{f} \mathrm{d}f \tag{2.1}$$

Where i and j are the lower and upper bound of the false positive fraction and were set to 0.000001 to 0.1, respectively, and $s(f)$ is the sensitivity at the false positive fraction f.

Statistical comparison was performed by means of bootstrapping[90]. On the test set, average ROC curves were calculated over 1000 bootstraps. Additionally, the mean sensitivity was calculated for each bootstrap and p-values were computed for testing significance[113,114]. Differences were considered to be significant for p-values $<$0.05.

2.3 Results

The mean sensitivity obtained from the ROC analysis are shown in Table 2.4. In this table, the mean sensitivity is shown for each individual vendor as well as on the whole test set. For each dataset, the CNN obtained a higher mean sensitivity for all datasets. On the complete test set, containing Hologic, GE, and Siemens data, the CNN achieved a significantly higher mean sensitivity compared to the cascade, 0.6914$\pm$0.0041 (mean$\pm$stdev) versus 0.6381$\pm$0.0038 (p$<$0.001). ROC curves were calculated on the whole test set for both methods and are shown in Figure 2.1. The ROC curves are plotted on a logarithmic scale to show the difference between the two methods at high specificity. In Figure 2.1, it can be seen that the CNN yields a higher sensitivity over the whole specificity range. All positive samples were detected at a specificity of 0.71 by the CNN while the cascade detects all positive samples at a specificity of 0.02. At a false positive fraction of 0.1, 0.01, 0.001, and 0.0001, the CNN detected 99.92%, 99.58%, 95.17%, and 74.63% of the positive samples, respectively. At the same false positive fractions, the cascade classifier detected 98,90% 95.79%, 90.85%, and 63.89% of the positives, respectively.

2.4 Discussion

Automated computer aided detection systems of calcifications in mammography have the potential to aid radiologists in reading mammograms. These systems are

Mean sensitivities			
Dataset	CNN	Cascade	p-value
Hologic	0.7035±0.0068	0.6534±0.0062	<0.001†
GE	0.7015±0.0069	0.6499±0.0069	<0.001†
Siemens	0.6726±0.0069	0.6180±0.0064	<0.001†
Hologic + GE + Siemens	0.6914±0.0041	0.6381±0.0038	<0.001†

Table 2.4: Mean sensitivities in the false positive fraction range of 0.000001 and 0.1 for the individual datasets. † Results are significantly different between the CNN and the cascade classifier ($p<0.05$).

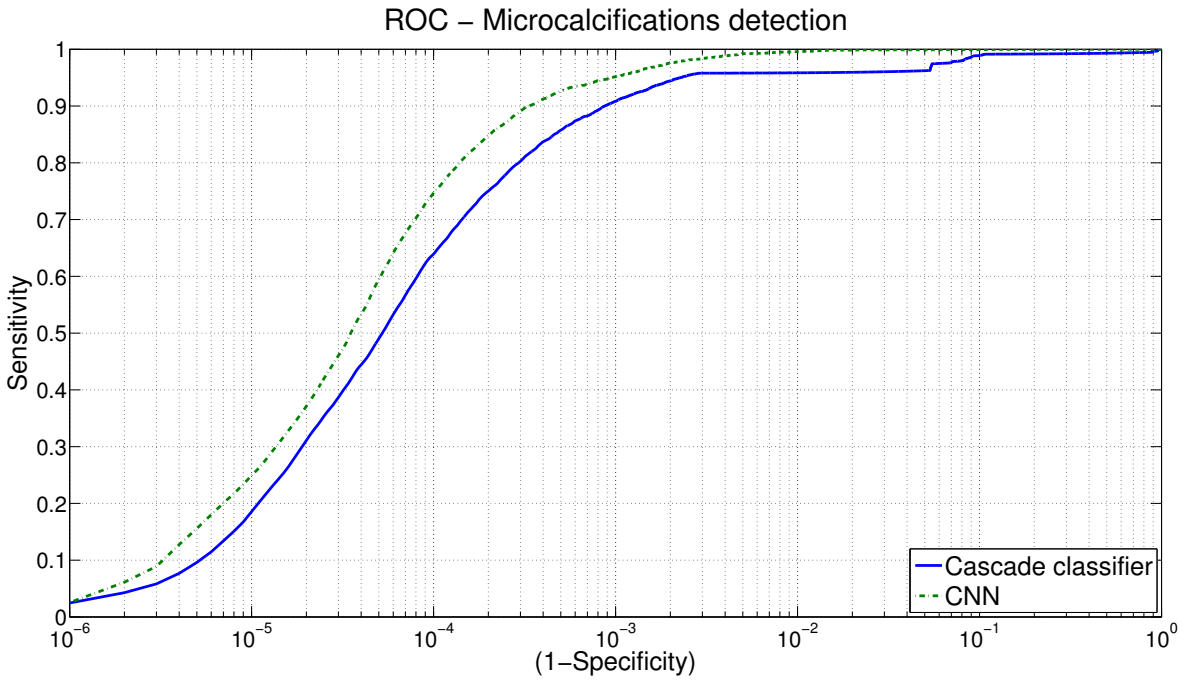


Figure 2.1: Average ROC curves of the calcification detection with the cascade classifier and the convolutional neural network of 1,000 bootstraps. The ROC curves are plotted on a logarithmic scale.

often designed into two stages: (1) detection of calcification candidates in the whole mammogram and (2) classifying calcification groups into benign and malignant. In this study, we focused on the detection of calcification candidates and we implemented a convolutional neural network for this task. To our knowledge, this is the first study where a deep learning strategy is developed for this task. The convolutional neural network was compared to a current state-of-the-art method, the cascade classifier. The comparison showed that the CNN outperforms the cascade classifier in terms of sensitivity in the whole specificity range. Additionally, the mean sensitivity in a false positive fraction range of 0.000001 to 0.1 was significantly higher with the CNN in the classification of all datasets acquired with three different mammography unit manufacturers.

Automatic selection of women with breast arterial calcifications

3

Jan-Jurre Mordang, Jakob Hauth, Gerard J. den Heeten, and Nico Karssemeijer

Original title: Automated labeling of screening mammograms with arterial calcifications

Published in: Breast Imaging, 2014

Abstract

For the automatic detection of malignant calcification clusters in screening mammograms a computer aided detection (CADe) system has been developed. The most frequent false positives of this system are breast arterial calcifications (BACs). The purpose of this study was to construct a method for selecting cases with BACs in mammographic screening data as part of a procedure to reduce false positives of the CADe system. To automatically select cases containing BACs, a GentleBoost classifier was trained. For composing the training set, the CADe system was applied on 10,000 normal cases. From these cases, 400 cases with the most significant false positives were included in the training set and an additional 200 cases with less obvious false positives. For testing, an independent test set was created by cluster detection of 1,000 normal cases and 95 malignant cases. After cluster detection 342 normal cases contained false positives and in 93 malignant cases true positive clusters were detected. In the training set, 244 cases showed signs of BACs and in the test set 95 cases. A total of 102 case-based features were calculated to train the classifier. A ROC curve was calculated of the classification of the test set bootstrapped 5000 times. The area under the curve of the ROC was 0.92 and already 44% of the cases with BACs were detected without any false positives. Furthermore, 90% of the cases with BACs were detected at a false positive rate of 20%. The performance of the proposed selection method implies a good feasibility to classify cases with BACs at high specificity. By using this selection we will be able to apply dedicated methods for false positive reduction due to BACs.

3.1 Introduction

Calcification clusters in the breast are a biomarker for breast cancer. For the purpose of automatic detection of these malignant clusters a computer aided detection (CADe) system has been developed. However, not all calcifications in the breast are malignant as benign calcifications can be observed in mammograms as well. The most frequent benign calcification clusters that are marked with a high malignancy likelihood by the current CADe system are breast arterial calcifications (BACs) (example shown in Figure 3.1). Therefore, by adding an additional false positive removal classification to our CADe system, specified on detecting BACs, might improve the system. Selecting cases with BACs and removal of false positive clusters in only these cases can prevent the removal of malignant clusters and ultimately lead to a more specific system for the detection of malignant calcification clusters. Therefore, the purpose of this study was to construct a method for selecting cases with BACs in mammographic screening data.

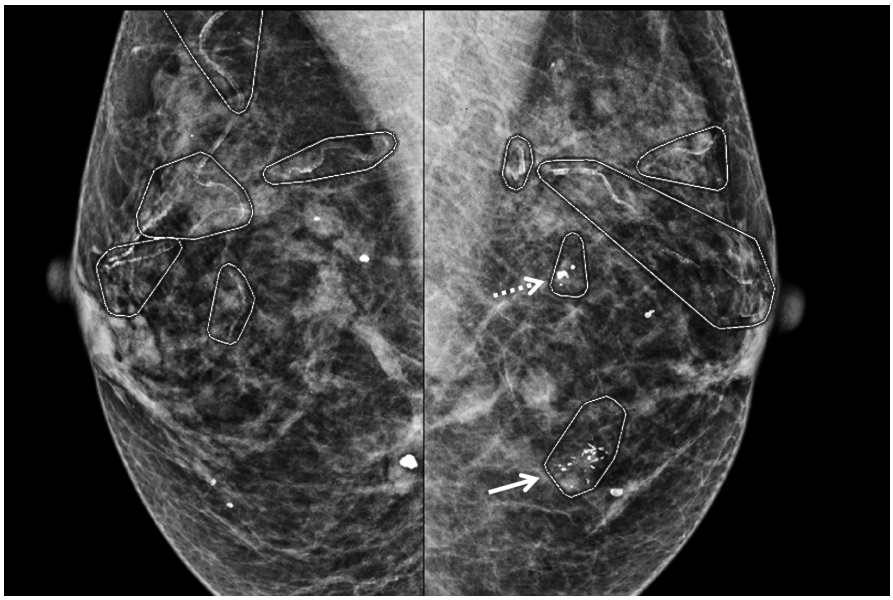


Figure 3.1: Example of calcification cluster detection in 2 mammograms containing BACs, macrocalcifications and a malignant calcification cluster. The malignant cluster is denoted with the solid arrow. A false positive due to a macrocalcification is denoted with the dashed arrow. All other annotations are false positives due to BACs.

3.2 Methods

The framework for the selection of cases with BACs consists of four stages. The first stage is the selection of the calcification candidates in raw screening mammograms

with a cascade classification scheme. In the next stage, the selected calcification candidates are clustered. And in the third stage, false positive clusters are removed with a trained classifier. The last stage consists of a case-based approach for the selection of cases with BACs. The first three stages are based on the work of Bria et al[74]. Therefore, these stages will only be touched very briefly in the next section. The fourth stage will be discussed in more detail in the subsequent section. A full flowchart of the framework is visualized in Figure 3.2.

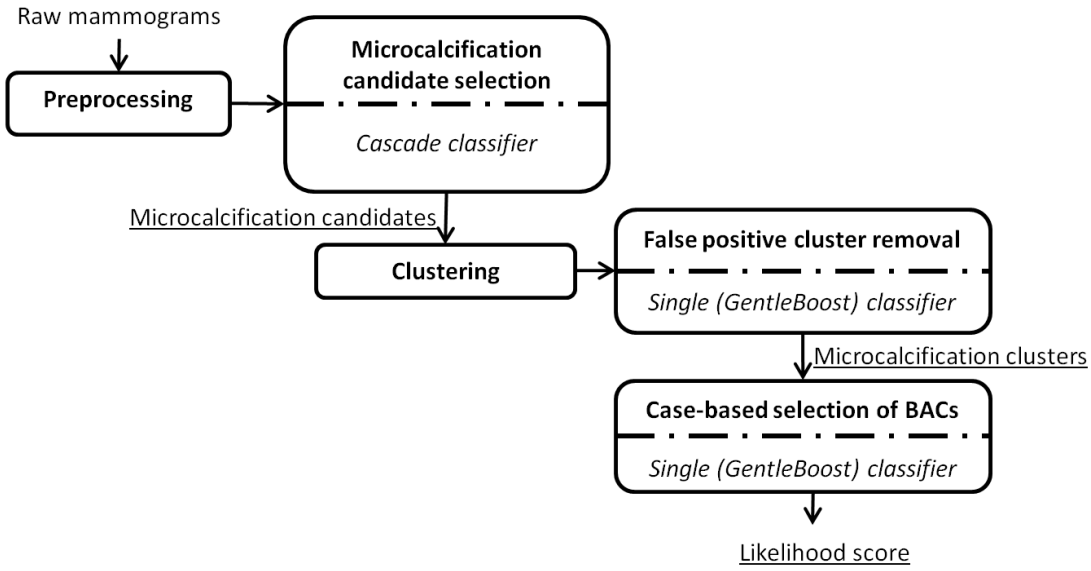


Figure 3.2: Flowchart of the whole framework. There are 4 main stages: i) calcification candidate selection, ii) clustering, iii) false positive cluster removal, and iv) classification of cases with BACs. Type of classification is denoted in *italic* and (intermediate) results are underlined.

3.2.1 Calcification cluster detection

For the detection of calcifications in raw mammograms, a cascade classifier is trained[77]. After preprocessing,for each pixel in the mammogram a patch is made with a dimensions of 13 x 13 pixels where the pixel lies in the center of the patch. This patch goes through 4 stages where in each stage the patch is classified by a GentleBoost classifier[78]. Features for each stage are determined during training from a total of 8 groups of Haar-like features[79]. These feature groups are be scaled and translated within the patch. Examples of these groups are shown in Figure 3.3. Patches classified as negative in one of the first three stages are removed and the remaining patches obtain a probability score in the last classification stage. For the four stages 3, 6, 11, and 51 haar-like features were calculated, respectively.

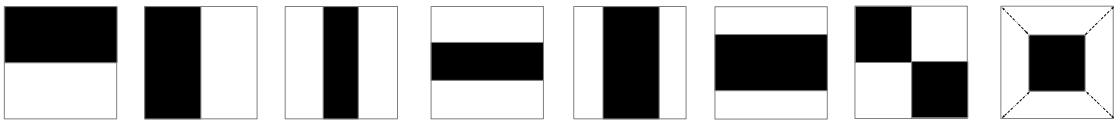


Figure 3.3: Examples of the haar-like features groups.

Calcification candidate classification in a mammogram leads to an image where each pixel corresponds to a probability score or zero if the patch is removed in an early stage of the cascade classifier. In these probability images, calcifications are segmented with connected component analysis. Macrocalcifications, calcifications larger than $1mm$, are removed as well as calcifications that lie within 2 pixels from a macrocalcification. Furthermore, a calcification is kept when at least 2 pixels have a probability above a preset threshold. Clusters are made of calcifications within a distance of $10mm$ to each other. Clusters containing less than 3 calcifications are discarded.

To remove false positives, a GentleBoost classifier is trained. For each cluster, features are calculated on the calcifications within the cluster and the cluster itself. These features ware based on shape, topology, probability, and texture. The GentleBoost classifier was trained on 100 regression stumps. As a result each detected cluster obtains a likelihood score.

3.2.2 Selection of cases with BACs

To determine cases with BACs, a multi-view classification procedure is carried out. In this procedure, all views of a case are analyzed, i.e. medio-lateral oblique and cranio-caudal views of the right and left breast. In these views, the calcification cluster detection is performed resulting in detected clusters with a likelihood score. For the multi-view analysis only likelihood scores above a specified threshold are considered. This threshold is set a the highest case-based sensitivity for malignant cases in the calcification cluster detection (100% detection rate at 33% false positive rate).

Features are calculated on a case level. These features are based on shape, topology, probability, texture, and vesselness[115]. For each cluster in each view a total of 24 cluster features are calculated. For each case (containing 2 or 4 views) the mean, standard deviation, maximum and minimum of all cluster features in each view are taken. Additionally, 6 case-based features are calculated based on the number of clusters in each view and the number of views. This leads to a total of 102 features per case. Table 3.1 shows the whole list of features for training of the classifier. On these features a GentleBoost classifier is trained using 50 regression stumps as weak

learners. The output of the classifier is a probability score for the presence of BACs in the case.

3.2.3 Performance evaluation

Several datasets were obtained from the Dutch Breast Cancer Screening Program (Bevolkings Onderzoek Midden-West, The Netherlands). For the calcification candidate selection and cluster detection, 2 datasets were composed. One dataset where individual calcification centers were annotated containing 129 abnormal cases (70 benign and 59 malignant). The second dataset contained cases where the contour was annotated of calcification clusters. This set included 186 abnormal cases (134 benign and 52 malignant) and 315 normal cases. The first dataset was used for training the classifier for calcification selection and the second dataset for training of the cluster classifier.

Two datasets were composed for the selection of cases with BACs. For the training set, cluster classification was carried out on 10,000 normal cases. From these cases a group of 400 normal cases with the most significant false positives and a group of 200 normal cases with less obvious false positives were included. In this training set, a researcher experienced in reading mammograms labeled each case if it contained BACs. The test set consisted of 1,000 normal and 95 malignant cases. In this set, cases with BACs were labeled by a resident of the radiology department. The normal cases in the training and test set were randomly selected from a database containing over 50,000 normal cases.

To evaluate the performance of the selection of cases with BACs, the trained classifier, trained on the training set, was tested on the test set. After classification, each case obtained a probability score. Of the classified dataset a Receiver Operating Characteristic (ROC) curve was made. The sensitivity is calculated by determining the number of cases with BACs labeled as positive divided by the total number of cases with BACs. The specificity is calculated by dividing the number of cases without BACs labeled as negative by the total number cases without BACs. The ROC curve was generated by bootstrapping the test set 5000 times.

3.3 Results

In the training set, 208 of the 400 cases with the most significant false positives showed signs of BACs and 36 cases in 200 cases with less obvious false positives. The test set contained 10 malignant and 98 normal cases with BACs. And after cluster detection, 342 normal cases were left over in the test set of which 87 cases contained

Cluster features	
cls Area	The area of the cluster.
cls Eccentricity	$\frac{I_{xx}+I_{yy}-\sqrt{(I_{xx}-I_{yy})^2+4I_{xy}^2}}{I_{xx}+I_{yy}+\sqrt{(I_{xx}-I_{yy})^2+4I_{xy}^2}}$ where I_{xx}, I_{yy} and I_{xy} are the moments of inertia.
cls Ellipse	The ratio between the long axis and the short axis of a fitted ellipse.
cls Number	The number of calcifications in the cluster.
cls Coverage	$\sum_{i=1}^{n} \frac{A_{mC_i}}{A_c ls}$ where A_{mC_i} is the area of the calcification i, n the number of calcifications within A_{cls}, the cluster area.
cls Density	$\frac{2\|E\|}{n(n-1)}$ where E is the number of edges of the graph.
cls Orientation	The orientation of the cluster with respect to the xy-plane.
cls Distance to skin/air	The distance of the center of the cluster to the skin air boundary.
cls Probability	Cluster probability from the cluster detection.
cls Hessian (5)	The Hessian-based vesselness filtered image at varying scale ($0.2 \leq \sigma \leq 1.0$, steps of 0.2)
cls Tubeness (5)	$k_{line}(\lambda_1, \lambda_2) = \frac{\lambda_2 - \lambda_1}{\lambda_2}$ where $\lambda_1 \leq \lambda_2$, the absolute eigenvalues calculated at varying scale ($0.2 \leq \sigma \leq 1.0$, steps of 0.2)
cls Lambda (5)	The highest absolute eigenvalue λ_2 at varying scale ($0.2 \leq \sigma \leq 1.0$, steps of 0.2)
Case features	
Case total cls	The number of clusters in the case.
Case cls per view (4)	The number of clusters per view. (Mean, standard deviation, maximum and minimum)
Case number of views	The number of views.

Table 3.1: Features for classifier training for the case-based selection of cases with BACs.

BACs. From the 95 malignant cases in the test set 93 cases were detected of which 8 cases contained BACs.

Figure 7.1 shows the ROC curve of the selection of cases with BACs plotted with 95% confidence intervals. The area under the curve of the ROC was 0.92. Furthermore, these results show that a sensitivity 0.44 is reached with no false positives up to a sensitivity of 0.90 at a specificity of 0.80.

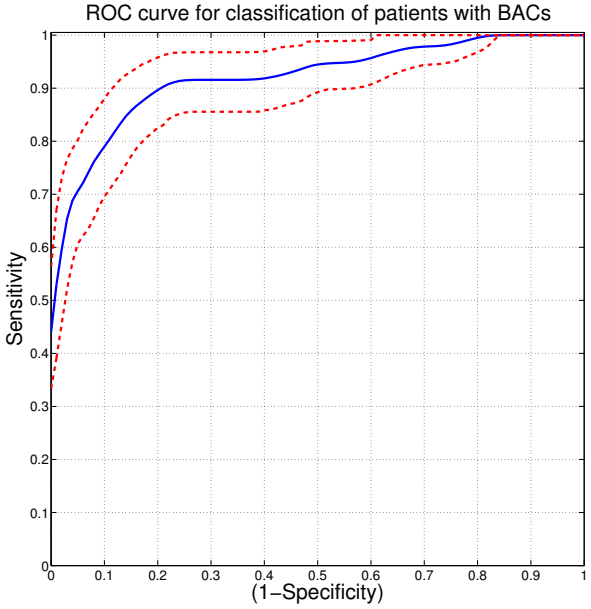


Figure 3.4: ROC curve of the classification of cases with BACs, bootstrapped 5000 times. 95% confidence intervals are plotted with the dashed lines.

3.4 Discussion

The percentage of cases with BACs found by the resident in the test set (9.7%) corresponds with the percentages found in literature[116–118]. Although BACs are of no interest in breast cancer screening, the presence of BACs is associated with atherosclerosis and cardiovascular disease[19,22,119]. Selection of these cases with the proposed method can also be used for the detection of diseases other than breast cancer.

Analysis of the 400 selected cases with the most significant false positives in the training set resulted in 32 cases (8%) with true false positives (e.g. obvious detection errors), 109 cases with calcifications (27%), 51 cases with macrocalcifications (13%), and 208 cases with BACs (52%). Showing that BACs are the most frequent false positives in our CADe system.

Several studies are done on automatic detection of vascular calcifications in the breast[120,121]. However, these studies are evaluated on the individually detected BAC

clusters. While the proposed method is based on the case-based selection method. This makes it difficult to compare the different methods. Nonetheless, false positive reduction in the cases with BACs, selected by the proposed system, still has to be done as a future work. An example of the flowchart for future work is shown in Figure 3.5.

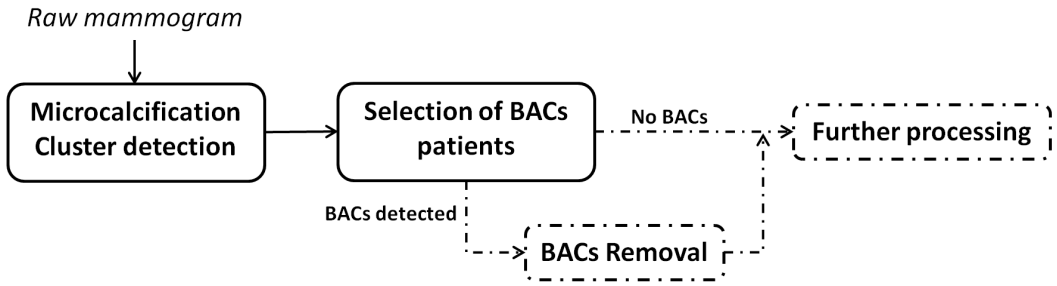
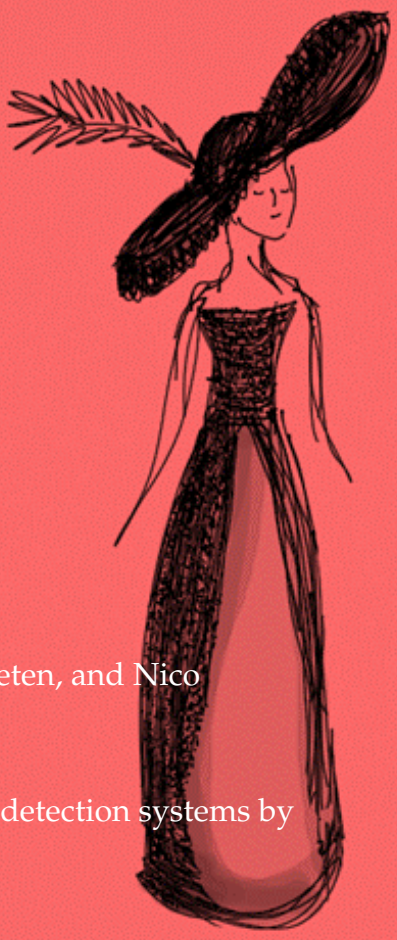


Figure 3.5: Flowchart of the framework for future work. The solid blocks and arrows are proposed in this study, The dashed blocks and arrows will be done in future studies.

The proposed framework shows a good performance for the selection of cases with breast arterial calcifications. By using this selection we will be able to apply dedicated methods for false positive reduction due to BACs while minimizing the risk of removing relevant true positive calcification clusters.

Removal of breast arterial calcifications as CAD findings in mammograms

4



Jan-Jurre Mordang, Albert Gubern-Mérida, Gerard den Heeten, and Nico Karssemeijer

Original title: Reducing false positives of microcalcification detection systems by removal of breast arterial calcifications

Published in: Medical Physics, 2016

Abstract

Purpose: In the past decades, Computer-Aided Detection (CADe) systems have been developed to aid screening radiologists in the detection of malignant calcifications. These systems are useful to avoid perceptual oversights and can increase the radiologists' detection rate. However, due to the high number of false positives marked by these CADe systems, they are not yet suitable as an independent reader. Breast Arterial Calcifications (BACs) are one of the most frequent false positives marked by CADe systems. In this study, a method is proposed for the elimination of BACs as positive findings. Removal of these false positives will increase the performance of the CADe system in finding malignant calcifications.

Methods: A multistage method is proposed for the removal of BAC findings. The first stage consists of a calcification candidate selection, segmentation and grouping of the calcifications, and classification to remove obvious false positives. In the second stage, a case-based selection is applied where cases are selected which contain BACs. In the final stage, BACs are removed from the selected cases. The BACs removal stage consists of a GentleBoost classifier trained on calcification features describing their shape, topology, and texture. Additionally, novel features are introduced to discriminate BACs from other positive findings.

Results: The CADe system was evaluated with and without BACs removal. Here, both systems were applied on a validation set containing 1088 cases of which 95 cases contained malignant calcifications. After bootstrapping, FROC and ROC analysis was carried out. Performance between the two systems was compared at 0.98 and 0.95 specificity. At a specificity of 0.98, the sensitivity increased from 37% to 52% and the sensitivity increased from 62% up to 76% at a specificity of 0.95. Partial areas under the curve in the specificity range of 0.8 to 1.0 were significantly different between the system without BACs removal and the system with BACs removal, 0.129 ± 0.009 versus 0.144 ± 0.008 ($p<0.05$), respectively. Additionally, the sensitivity at one false positive per 50 cases and one false positive per 25 cases increased as well, 37% versus 51% ($p<0.05$) and 58% versus 67% ($p<0.05$) sensitivity, respectively. Additionally, the CADe system with BACs removal reduces the number of false positives per case by 29% on average. The same sensitivity at one false positive per 50 cases in the CADe system without BACs removal can be achieved at one false positive per 80 cases in the CADe system with BACs removal.

Conclusions: By using dedicated algorithms to detect and remove breast arterial calcifications the performance of CADe systems can be improved, in particular at false positive rates representative for operating points used in screening.

4.1 Introduction

Breast cancer is one of the leading types of cancer among women in terms of new cases and deaths[39,122]. Early detection of this disease is essential to decrease mortality[34,35]. Therefore, breast cancer screening programs are implemented in many countries to detect breast cancer at an early stage. This has proven to be a successful approach in a number of wealthy countries[39]. In mammography, which is the imaging modality used for screening, the presence of calcifications is a sign for Ductal Carcinoma-In-Situ (DCIS). Early detection of calcifications associated with DCIS is important because of the relative high risk that DCIS will develop into invasive breast cancer[32,33] which occurs in over 40% of nonpalpable breast cancer cases[36]. However, the detection of calcifications in mammograms and their characterization is a tedious task. Calcifications might be subtle and can be easily overlooked, while sometimes it is hard to determine whether calcifications are really present or that patterns similar to calcifications are simulated by noise or specific image processing algorithms. Furthermore, various types of calcifications can be present in the breast of which many represent benign disease[123]. This makes it difficult for screening radiologists to decide whether a woman should be recalled or not[66].

To aid screening radiologists in finding calcifications, and helping them to improve the positive predictive value of their recalls, Computer-Aided Detection (CADe) systems have been developed. These systems automatically analyze mammograms and mark locations which are suspicious for abnormalities. During screening, the radiologists' attention is drawn to these locations which prevents that abnormal regions are overlooked. In the past decades, much research has been done in the development of CADe systems and it is still a prominent research subject to this day[61,62,68–72]. Current CADe algorithms for calcification detection have a good sensitivity at a cost of around two false positive findings per screening mammogram. A normal screening mammogram consists of four mammographic images, a Cranial-Caudal (CC) and a Medial-Lateral Oblique (MLO) view of each breast. The benefit of using a CADe system in screening has been analyzed in several studies. It has been found that, when CADe is used in daily screening practice the detection rate of radiologists increases[73,94–96]. However, the recall rate increases as well due to the high number of false positive marks of the CADe system[73,97–100]. These studies show that, although CADe systems are useful to avoid perceptual oversights of the radiologist, they are not yet suitable to serve as an independent observer[66,124]. In order to consider the use of CADe as an independent reader, the recall rate of the system should be in the order of what is achieved in breast cancer screening programs. In Europe, for example, the recall rate is less than 5% and in the United States it is

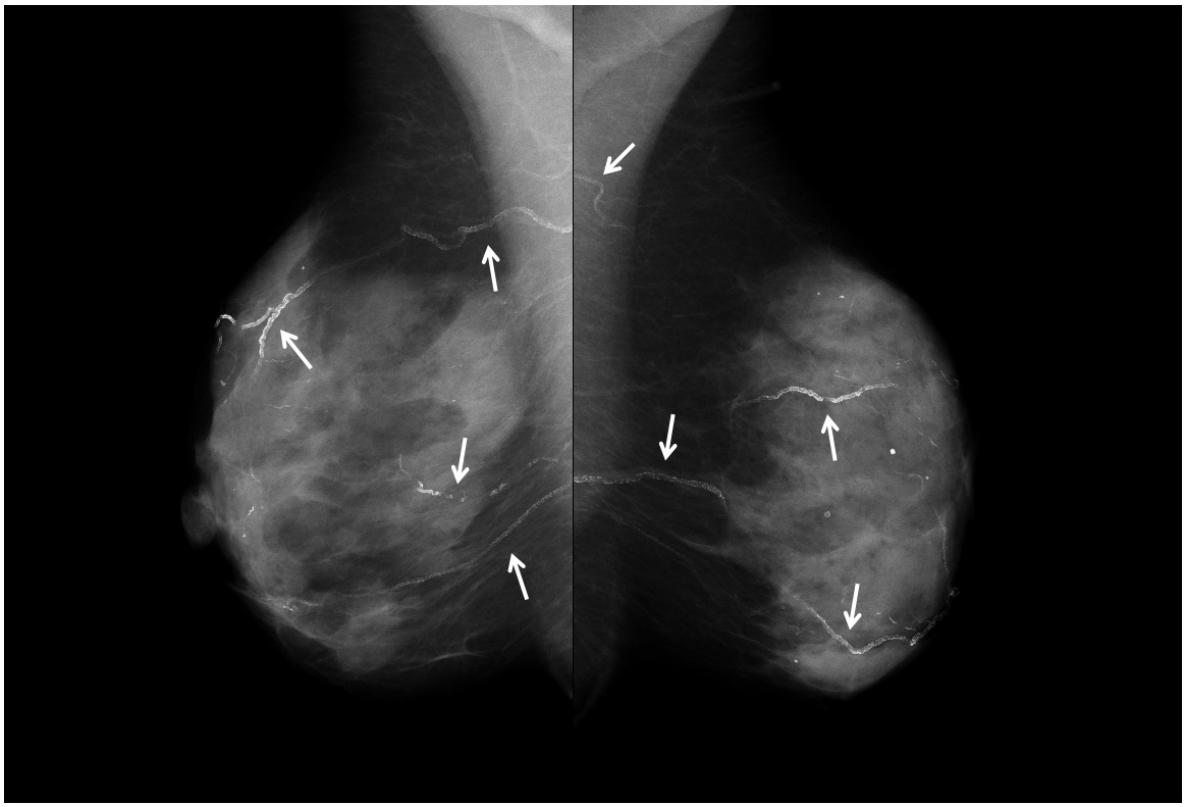


Figure 4.1: An example of a case with BACs. BACs can be observed in both views and are denoted by the white arrows.

around 10%. Approximately a third of the recalled cases contains calcifications[125,126]. These numbers indicate that the number of false positive marks of CADe systems should decrease by one or two orders of magnitude.

Most false positive findings in the detection of calcifications arise from various types of benign calcifications in the breast. One of the most frequent false positives detected by a CADe system originates from Breast Arterial Calcifications (BACs)[127]. These benign calcifications are small calcium deposits in the vessel wall of the arteries. An example of a case containing BACs is shown in Figure 4.1. Although they may occasionally resemble intraductal calcifications, for a screening radiologist BACs are generally easily dismissed as non-suspicious calcifications. For a CADe system to dismiss BACs is more difficult and many of them are still detected. Removal of BACs as positive findings is essential in order to increase the performance of automated detection systems. The aim of this paper is to contribute to the development of a CADe system which performs as good as or better than a screening radiologist. This means that the system has to operate at a very high specificity (e.g. one false positive per 50 cases) while maintaining a high sensitivity for finding malignant calcifications.

The main purpose of this study was to develop and validate a CADe system in

which BACs are detected and removed as positive findings. We have chosen to remove BACs after the detection of calcifications. An alternative approach would be to remove vascular structures with a curvilinear detector before the calcification detection[128]. This could be done by applying a vessel mask to exclude vascular regions. However, segmentation of vascular structures is a challenging task because vessels appear as non-continuous structures throughout the whole breast. In fact, often they can only be recognized visually due to presence of BACs. Therefore, we did not follow this approach. Instead, we aimed at classifying CADe findings into BACs or malignant findings using a supervised learning strategy where classifiers were trained on a dataset containing 900 cases. In the literature, several classifiers have been used for this task such the linear discriminant classifier[129], the Support Vector Machine classifier[130,131], artificial neural networks[132], and several boosting classifiers such as AdaBoost, RankBoost, and GentleBoost[74,133]. Due to its fast optimization and its resistance to overfitting, GentleBoost classifiers are used in this study. For validation, a large independent set with 1088 cases was used, of which 95 cases contain malignant calcifications.

4.2 Methods

4.2.1 Calcification detection

The CADe system that is under development in this paper is based on the *CasCADe* system proposed by Bria et al[74]. This system uses a multistage method for the detection of calcifications and can be applied to Full Field Digital Mammograms (FFDMs). In this paper, we will only discuss this system very briefly. The calcification detection system consists of three stages, i) calcification candidate selection, ii) calcification segmentation and grouping, and iii) calcification group classification. From this point, the calcification detection system will be referred to as the initial CADe system. Before the calcification candidate selection, a noise equalization method[134,135] is applied to each mammogram and the breast is segmented from the background.

Calcification candidate selection

In this stage, each pixel in the segmented breast tissue region in a mammographic image is classified with a cascade classifier[77]. The cascade classifier consists of several nodes where in each node a GentleBoost[78] classifier is trained. Each GentleBoost classifier is trained on regression stumps[78] and Haar-like[79,80] features are used for classification. To classify each pixel in the image, a 13x13 pixel patch is extracted for all pixels. On each patch, various Haar-like features groups are calculated with

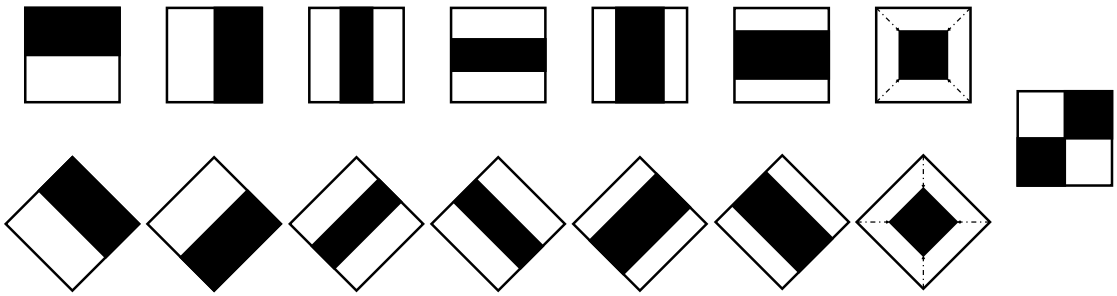


Figure 4.2: Haar-like feature groups, each feature group consists of features at all possible scales and translations within a 13x13 pixel patch window.

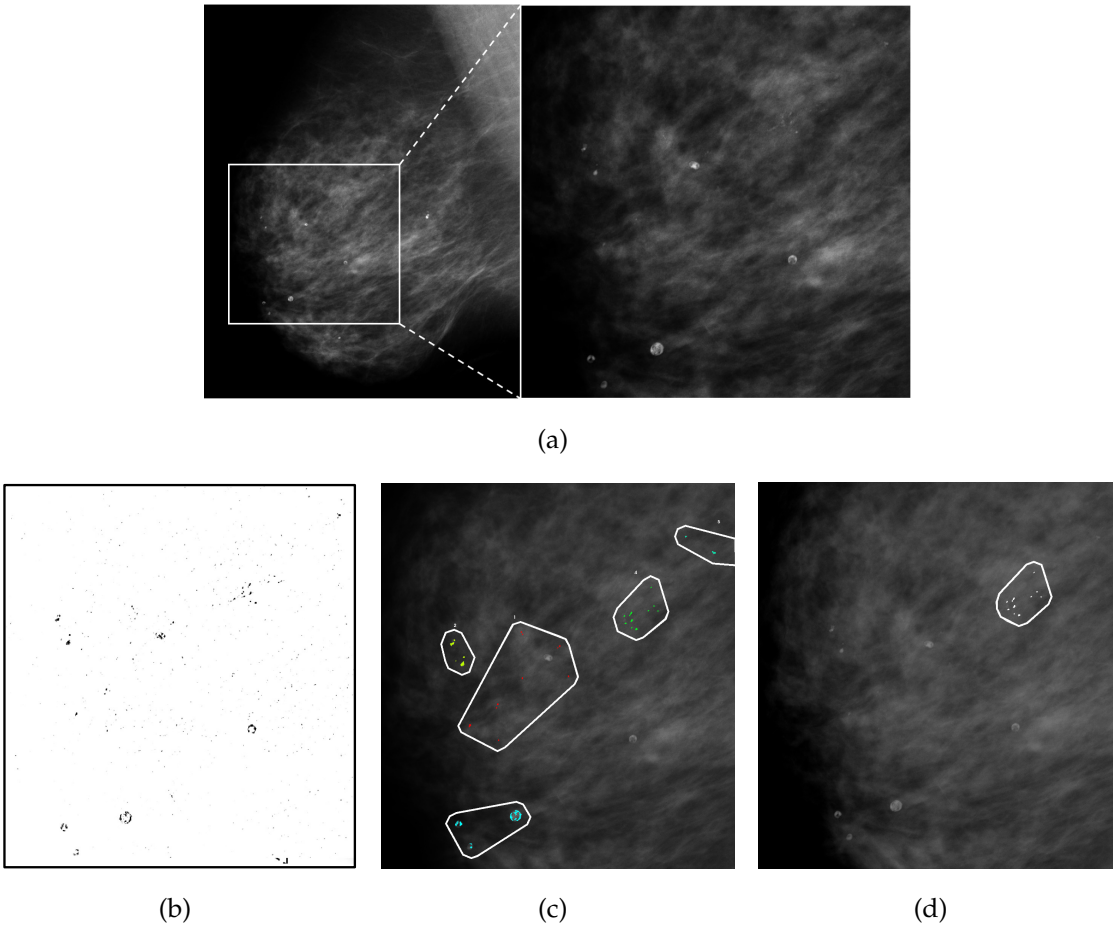
various scales and at various locations within the patch. Examples of the Haar-like feature groups are shown in Figure 4.2. In every node of the cascade, each pixel is classified and pixels with a classification score below a certain threshold are removed for further classification in the subsequent nodes. In the last node, each remaining pixel, i.e. a pixel classified as a positive pixel in all nodes, receives a final classification score. This score is the output of the last classifier. All other pixels receive the value zero. An example of the output of the classifier is shown in Figure 4.3(b).

Calcification segmentation and grouping

Calcifications are segmented by applying a connected-component analysis to all pixels with a final classification score obtained from the cascade. To remove macrocalcifications, components larger than $1mm$ are deleted. After segmentation, groups are formed by clustering calcifications which are located within $10mm$ from each other. Groups containing less than three calcifications are removed. In Figure 4.3(c), an example is shown of the detected calcification groups.

Group classification

A classification step is carried out to remove the most obvious false positive groups. In this classification, a single GentleBoost classifier, trained on 100 regression stumps, classifies all detected calcification groups. The features described in Bria et al[74] and Veldkamp et al[81] are used for this purpose. After classification, each detected calcification group receives a score reflecting how confident the system is that the group is suspicious. A threshold (T_{Group}) is applied on these scores to obtain the final result of the CADe system. Groups with a low confidence score, i.e. with a score below



(a)

(b) (c) (d)

Figure 4.3: Examples of intermediate results of the calcification detection; (a) the input screening mammographic image and a zoomed in region, (b) the calcification candidate selection, the black pixels represent locations of the calcification candidates, (c) the detected groups after calcification segmentation and grouping, and (d) the remaining calcification group after classification.

T_{Group}, are not considered for further analysis. An example of a detected calcification group after the group classification stage is shown in Figure 4.3(d).

4.2.2 Selection of cases with BACs

A method to select cases which have a high likelihood of containing BACs is used in this study. The selection stage is of importance because only a relatively low percentage (9 - 24%) of the western population shows signs of arterial calcifications[116–118]. By adding this stage we reduce the risk that malignant calcifications are erroneously removed in cases that do not contain any BACs. The selection method consists of a case-based classification process in which the whole mammogram of a case is considered. This approach is chosen because often the presence of BACs is more evident due to their appearance in multiple mammographic views, while individual BACs

may be hard to characterize as such. The selection method is previously proposed by Mordang et al[127].

For all available views of a single case, group features are calculated of all calcification groups with a score above T_{Group}. A total of 24 group features are calculated. These are based on shape, topology, probability, texture, and vesselness. Then, for each of the 24 features the mean, standard deviation, maximum and minimum are calculated of all groups and used as case features. Six additional case features are calculated based on the number of groups in all the images of a case and their distribution within the views. In the end, a total of 102 case features are used. Each case is classified by a GentleBoost classifier with 50 regression stumps and receives a score which represents the confidence of the classifier that a case contains BACs. To select cases with BACs, a threshold (T_{Case}) is set on the confidence scores.

4.2.3 Removal of BACs

This section focuses on the detection and removal of BACs. To do this, a false positive removal procedure is applied on only cases which are likely to have BACs, i.e. have a confidence score higher than T_{Case}. In each selected case, the detected calcification groups are classified by a GentleBoost classifier trained with 50 regression stumps. This classifier aims to discriminate between calcification groups of BACs and groups without BACs (non-BACs). The latter group includes malignant calcifications, benign calcifications, and other false positives detected by the system. To train the classifier, 14 feature types are calculated for each detected calcification group. These feature types, with a brief description, are shown in Table 4.1. Many of these features are already described in the literature[74,81]. Therefore, we will only discuss the novel features which are designed for the classification of BACs.

When radiologists determine if calcifications are located in the arteries, they look at their distribution and their location in a blood vessel. Because BACs are located in the arterial wall, their distribution will be elongated along the vessel. Therefore, novel features are proposed which are designed to distinguish the elongated distribution of BACs from the more concentrated distribution of other calcification types. An example of each of the two calcification distributions is shown in Figure 4.4. One of the novel features is the *elongatedness* of a calcification distribution. The *elongatedness* is defined as the ratio between the overall length and width of the calcification group. To determine the *elongatedness*, a piece-wise linear analysis is performed. Here, the calcification group is divided into subgroups with a radius of $1cm$. Then, eigenvectors are calculated of each subgroup with Principle Component Analysis (PCA). For each subgroup, the ratio between the two eigenvalues is com-

Group-based features	
Area	The area of the convex hull[136] fitted on the calcifications within the group.
Eccentricity	$\frac{I_{xx}+I_{yy}-\sqrt{(I_{xx}-I_{yy})^2+4I_{xy}^2}}{I_{xx}+I_{yy}+\sqrt{(I_{xx}-I_{yy})^2+4I_{xy}^2}}$ where I_{xx}, I_{yy} and I_{xy} are the moments of inertia.
Ellipse	The eccentricity of a fitted ellipse on the detected group. Ellipse is $\sqrt{1-ds^2/dl^2}$ where ds and dl are the short side and long side, respectively.
Number	The total number of calcifications in the group.
Coverage	$\sum_{i=1}^{n} \frac{A_{mC_i}}{A_c ls}$ where A_{mC_i} is the area of the calcification i, n the number of calcifications within A_{cls}, the group area.
Density[137]	$\frac{2\|E\|}{n(n-1)}$ where E is the number of edges of the graph.
Orientation	The orientation of the group with respect to the xy-plane.
Distance to skin/air	The distance of the center of the group to the skin air boundary.
Probability	Probability from the group classification.
Elongatedness	The ratio of the eigenvalues calculated with principle component analysis on subgroups with $1cm$ radius
Circle coverage	The ratio of the group surface area and the area of the minimum enclosing circle
Hessian (5 feature values)	The Hessian filtered image at varying scale ($0.2 \leq \sigma \leq 1.0$, steps of 0.2)
Tubeness (5)	$k_{line}(\lambda_1, \lambda_2) = \frac{\lambda_2-\lambda_1}{\lambda_2}$ where $\lambda_1 \leq \lambda_2$, the absolute eigenvalues calculated at varying scale ($0.2 \leq \sigma \leq 1.0$, steps of 0.2)
Lambda (5)	The highest absolute eigenvalue λ_2 at varying scale ($0.2 \leq \sigma \leq 1.0$, steps of 0.2)

Table 4.1: Proposed features for the detection of BACs.

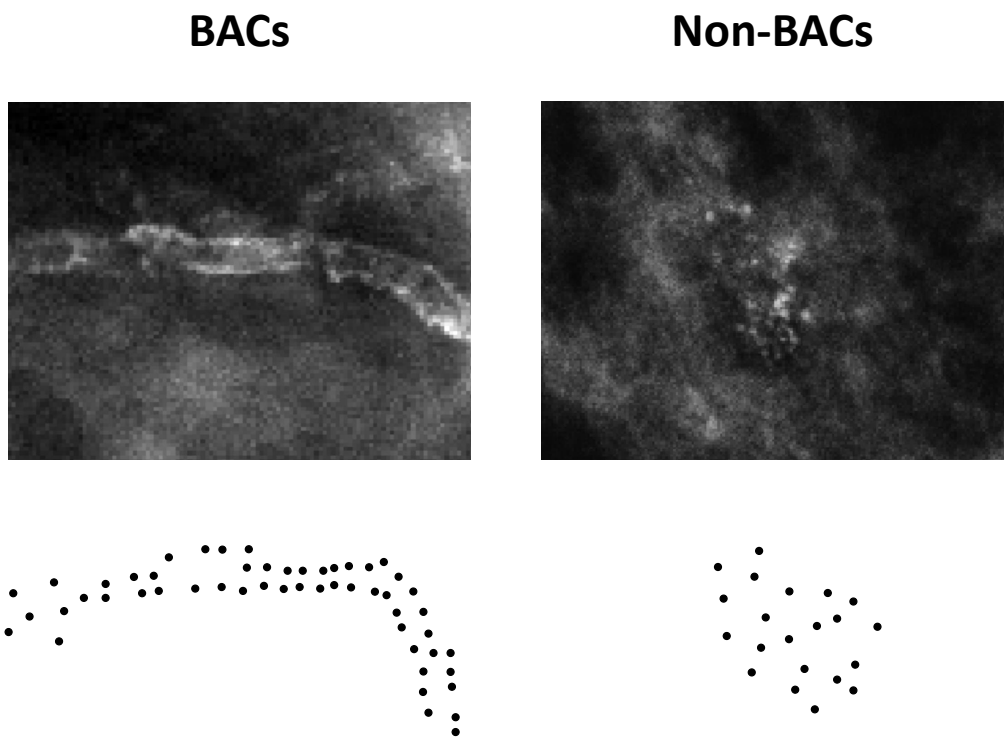


Figure 4.4: Example images of two types of calcification distributions and their schematic representation. On the left side, an example of BACs. On the right side, an example of non-vascular calcifications. Every dot in the schematic drawings represents the center of a calcification.

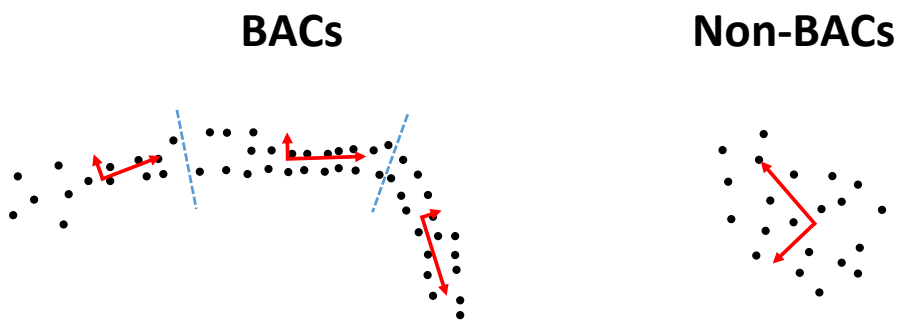


Figure 4.5: Schematic drawings of the elongatedness feature for a BAC and a non-BAC distribution. Each calcification group is split up into subgroups (denoted by the dotted lines). For each subgroup, the eigenvectors are calculated with PCA (denoted by the arrows). The elongatedness feature is the average ratio between the eigenvalues of all subgroups.

puted where the largest eigenvalue is divided by the smallest. The *elongatedness* is determined by calculating the average ratio of all subgroups. In Figure 4.5, an example is shown of the *elongatedness* feature. Because the shape of the BAC distribution is elongated, one eigenvalue is bigger than the other resulting in a very large ratio. A more concentrated calcification distribution will have almost equal eigenvalues and, consequently, have an *elongatedness* value close to 1.

One shortcoming of the *elongatedness* feature is that it fails when BACs are settled in a bifurcation of a vessel. Therefore, another feature is proposed which we call the circle coverage. This feature is defined as the ratio between the area covered by the calcifications and the area of a minimum enclosing circle fitted on the calcification distribution. To obtain the area covered by the calcifications, on the center of each calcification a disk is placed with a radius half of the distance to its closest neighboring calcification. The sum of all disks is taken as the coverage of the whole group. To reduce the influence of outliers, the mean radius of all disks is calculated and disks with a radius bigger than two times the standard deviation are ignored. The circle coverage is defined as the ratio between the area of the minimum enclosing circle and the coverage. An example is given in Figure 4.6.

Additionally to the two features describing the calcification distribution of BACs,

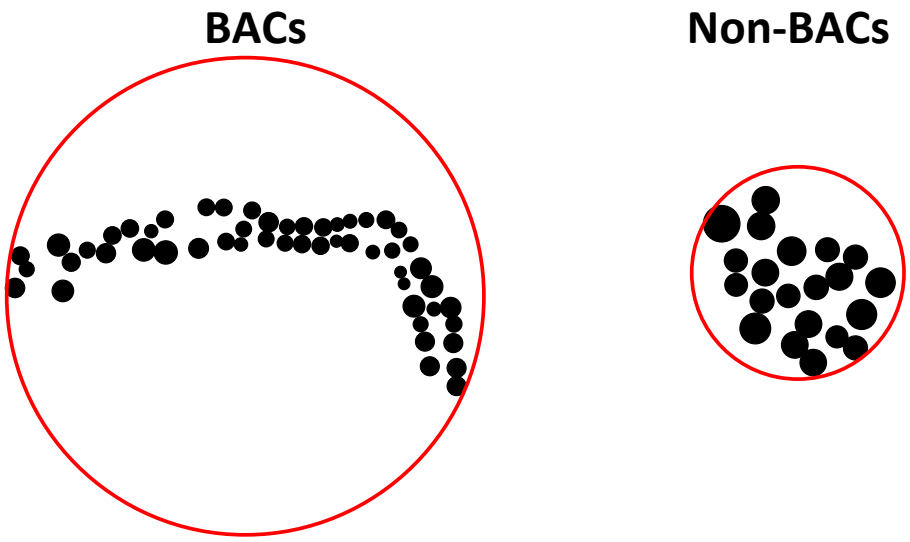


Figure 4.6: Schematic drawings of the circle coverage feature for a BAC and a non-BAC distribution. On the center of each calcification a disk is placed. A minimum enclosing circle is fitted on the calcification group (denoted by the circles). The circle coverage value is the ratio between the total area of all disks and the area of the minimum enclosing circle.

three types of features are calculated which represent the likelihood that a vessel is present at the location of the detected calcifications. These features are calculated with image filtering techniques to enhance tube- and vessel-like structures which are the Hessian, tubeness, and lambda filters. The *Hessian* feature is calculated on the mammographic images convoluted with a Hessian[138] filter. The *Hessian* feature is determined by calculating the average Hessian, which is the determinant of the Hessian matrix, of all pixel locations of the calcifications in the Hessian filtered image. The Hessian filter is applied with various scales ranging from $0.2 \leq \sigma \leq 1.0$ with steps of 0.2 resulting in a total of five feature values. The *tubeness* feature is calculated in a similar fashion. Here, the mammographic images are filtered with a tubeness[115] filter. The *tubeness* feature is calculated by averaging the tubeness value of all pixels locations of the calcifications in the tubeness filtered image. The *tubeness* feature is calculated at the same five scales as the Hessian feature resulting in five feature values. The third vascular feature is the *lambda* feature. On the Hessian filtered image an eigenanalysis is performed and for each pixel in the image the highest absolute eigenvalue (λ_2) is taken. To determine the *lambda* feature, the average λ_2 of all calcification locations is calculated. The *lambda* feature is calculated at the same five scales as the other two vascular features.

In summary, to eliminate BACs as positive findings, each detected calcification group with a score above T_{Group} in cases with a selection score above T_{Case} is classified. After classification, each group receives a likelihood score related to the confidence of the classifier for a group to contain BACs. To remove suspected BACs, all calcification groups above a specified threshold (T_{BACs}) are removed as positive findings.

4.2.4 Datasets

To train and validate the proposed system, three datasets are composed. 1) a dataset for training the initial CADe system, 2) a dataset for training of the case-based selection stage and the BACs removal stage, and 3) a dataset to validate the new system. There is no overlap of cases between the three datasets. All cases are selected from a large database consisting of over 50,000 cases obtained from the Dutch Breast Cancer Screening Program Database (Bevolkings Onderzoek Midden-West, The Netherlands). The mammographic images in this database are acquired with Hologic digital mammography systems (Hologic, Bedford, Massachusetts, United States) and have an isotropic pixel resolution of $70\mu m$. Most cases in the database include multiple screening exams. However, for all cases in this paper only one exam is included where each exam (or mammogram) consists of all available raw screening FFDM

images.

The first dataset contains 689 cases. 371 cases of these cases are recalled in screening and 318 are normal cases (cases which are not recalled). Biopsies or diagnostic follow up showed that 111 malignant calcification groups and 204 benign groups are present in the recalled cases. In 113 calcification groups, the centers of 7,787 calcifications are individually annotated. Contours are drawn for the remaining 202 calcification groups. All the annotations are made based on the diagnostic reports.

The second dataset contains 900 cases. From these cases, 300 cases are recalled in screening and 600 are normal cases. The normal cases are selected from a large pool containing 10,000 normal cases. Selection is done by classifying the large pool with the initial CADe system. Then, the detected calcification groups in these cases are ranked according to their CADe scores from high to low. From this ranking, the first 400 cases with the highest scores are selected. Additionally, 200 random cases which contained detected calcification groups with a CADe score above T_{Group} are selected and added to the dataset. Each detected calcification group in all 900 cases of the dataset is reviewed and labeled as BACs or non-BACs. Labeling is done by a researcher who is experienced in reading mammograms.

The third dataset, contains 1088 cases of which 95 cases contain malignant calcifications and 993 normal cases. A total of 196 malignant calcification groups are annotated in the malignant cases based on the diagnostic reports. Additionally, in all normal cases, all detected calcifications with a CADe score above T_{Group} were visually assessed and labeled as BACs or non-BACs.

4.2.5 Experiments and performance evaluation

The first data set was used to train the initial CADe system. This system was then applied on the cases in the second and third dataset resulting in detected calcification groups and their CADe scores. The calcification groups detected in the second dataset were used for training the case selection and the BACs removal stage. The calcifications detected in the third dataset were used for validation of the proposed framework.

For training and testing, T_{Group} was fixed throughout all the experiments. T_{Group} was set to the value that achieved, at maximum specificity, a sensitivity of 100%. To determine the value of T_{Group}, a 10 fold cross-validation on the dataset used for training the initial CADe system was performed. The values of T_{Case} and T_{BACs} were explored by means of a grid search where both thresholds were varied in the range of 0.1 and 0.9 with steps of 0.1. The combination of threshold values that yielded to the highest pAUC on the validation set was chosen. These thresholds were fixed for

training and testing the proposed system.

For evaluation, the CADe system with BACs removal is compared to the initial CADe system. To do this, all the cases in the validation set are processed by both CADe systems. After processing, each detected calcification group contains a CADe score. This score reflects the confidence of the CADe system for a group to be suspicious. On the CADe scores, a case-based Receiver Operating Characteristics (ROC) analysis is performed as well as a Free-response ROC (FROC) analysis. For the analysis, from each malignant case only the true positive with the highest score is taken where a true positive is defined as a detected calcification group with at least two calcifications located within an annotated malignant lesion. Malignant cases without any true positive findings are defined as false negatives. From each normal case, the detected calcification group with the highest CADe score is taken as a false positive.

To compare both systems, average ROC curves are created with bootstrapping[90]. Here, the average ROC curve is calculated after bootstrapping the validation set 5,000 times. Furthermore, the partial Area Under the Curve (pAUC) is calculated of each bootstrap where the pAUCs in the specificity range of 0.8 to 1.0 are statistically compared[113,114]. Average FROC curves are also created with bootstrapping the validation set 5,000 times. To statistically compare the two CADe systems, the sensitivity at one false positive per 50 cases and one false positive per 25 cases is calculated for each bootstrap and compared between the two systems. Additionally, the influence of the case-selection is assessed by computing the ROC curve of the CADe system with BACs removal but without the case-based selection, i.e. by setting T_{Case} to 0.0. The area under the ROC curve is statistically compared to the other two systems with bootstrapping. Additionally, the influence of the case-selection is assessed by computing the ROC curve of the CADe system with BACs removal but without the case-based selection, i.e. by setting T_{Case} to 0.0. The area under the ROC curve is statistically compared to the other two systems with bootstrapping.

4.3 Results

All cases in the validation set were classified by the initial CADe system and the CADe system with BACs removal. In total, 11 groups of malignant calcifications were missed by the initial CADe system which lead to two missed malignant cases. The CADe system with BACs removal missed one additional group. However, this did not lead to a missed case because this malignant calcification group was found in the other view. Labeling of the 900 cases in the training set resulted in 293 cases which contained BACs. In these cases a total of 729 calcification groups were labeled

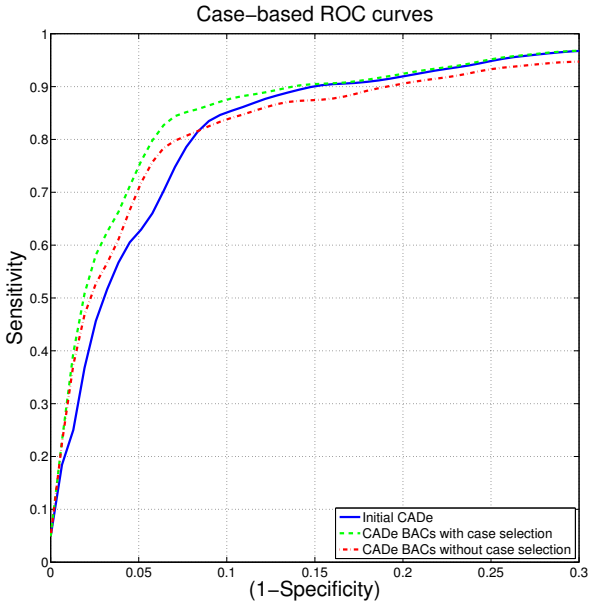


Figure 4.7: Case-based ROC curves of the initial CADe system, the CADe system with BACs removal and case-based selection and the CADe system with BACs removal without the case-based selection. The ROC curve is plotted in the range of 0.7-1.0 specificity. All curves are bootstrapped 5000 times and the average curves are shown.

as BACs. In the cases without BACs, 2159 groups were labeled as non-BACs. In the validation set, 542 calcifications groups were detected in 239 cases. Of these 542 calcification groups, 215 groups in 74 cases were visually assessed as BACs. The BACs removal approach removed 70 BACs groups in 35 cases.

In Figure 4.7, the ROC curves obtained with the initial CADe system, the proposed CADe system with BACs removal and the CADe system without the case-based selection are shown. The ROC curves show that in the range of specificity between 0.8 and 1.0, more malignant calcification groups are detected by the CADe system with BACs removal. At a specificity of 0.98, 37% of the malignant lesions were detected by the initial CADe system while 52% of the lesions were detected by the CADe system with BACs removal. Moreover, at a specificity of 0.95, more malignant lesions were detected by the CADe system with BACs removal than by the initial CADe system, 76% versus 62%, respectively. In Table 4.2, the comparison of the pAUC values are shown. This table shows that the pAUC of the CADe system with BACs removal is significantly higher with and without the case-based selection than the pAUC of the initial CADe system, 0.144 ± 0.008 and 0.141 ± 0.008 versus 0.129 ± 0.009 ($p<0.00002$ and 0.0006), respectively.

In Figure 4.8, the FROC curves are shown for the initial CADe system and the CADe system with BACs removal. These curves show that more malignant calcifications were detected at 0.2 False Positives per Case (FP/C) or fewer by the CADe

pAUC comparison			
	Mean	SD	p-value
Initial CADe	0.129	0.009	
CADe BACs with case selection	0.144	0.008	$<0.00002^{*}$
CADe BACs without case selection	0.141	0.008	0.0006^{*}

Table 4.2: Mean pAUC values of the case-based ROC analysis for the initial CADe system and the CADe system with BACs removal. The pAUC was calculated in the range of 0.8-1.0 specificity. * Results are significantly different from the initial CADe system $p<0.05$.

Sensitivity comparison						
	one FP per 50 cases			one FP per 25 cases		
	Mean	SD	p-value	Mean	SD	p-value
Initial CADe	0.37	0.09		0.58	0.07	
CADe with BACs removal	0.51	0.08	0.0068^{*}	0.67	0.07	0.029^{*}

Table 4.3: Mean sensitivity values at two operating points obtained with the FROC analysis and bootstrapping. The sensitivity is calculated for one false positive per 50 cases (0.02 FP/C) and one false positive per 25 cases (0.04 FP/C). * Sensitivity is significantly different between the initial CADe system and the proposed system, $p<0.05$.

system with BACs removal compared to the initial CADe system. In Table 4.3, the sensitivity of both systems is compared at one false positive per 50 cases (0.02 FP/C) and at one false positive per 25 cases (0.04 FP/C). At one false positive per 50 cases, the sensitivity of the CADe system with BACs removal was significantly higher than the sensitivity of the initial CADe system, 51% versus 37%, respectively. This can also be observed at one false positive per 25 cases where 57% of the malignant cases were detected by the initial CADe system and 67% by the CADe system with BACs removal. Furthermore, at a sensitivity of 1 false positive per 50 cases (0.02 FP/C) with the initial CADe system, the CADe system with BACs removal achieves the same sensitivity at 1 false positive per 80 cases (0.0125 FP/C). In the range of a sensitivity of 10% to 90%, the average difference in FP/C between the two curves is 29% with a peak of 44% at a sensitivity of 25%.

The influence of the case-based selection is shown in Figure 4.7. The CADe

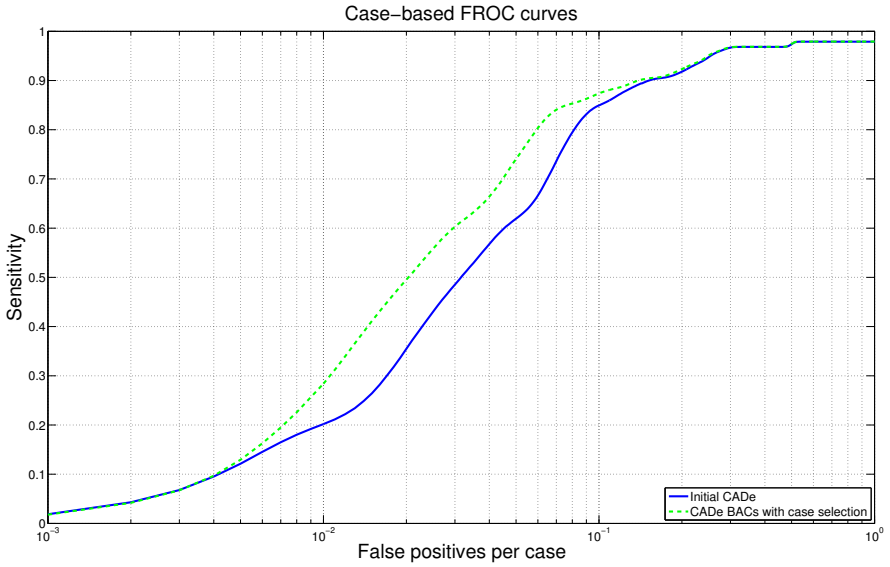


Figure 4.8: Case-based FROC curves of the initial CADe system and the CADe system with BACs removal. The FROC curve is plotted logarithmically from 1 false positive per 1,000 cases to 1 false positive per case. The curve is bootstrapped 5,000 times and the average curves are shown.

system without the case-based selection removed 12 additional malignant calcification groups leading to a total of 23 missed malignant calcification groups. Consequently, two additional malignant cases were missed with this system which led to a maximum sensitivity lower than the initial CADe system. In Figure 4.9, four missed malignant calcification groups that were missed by the CADe system without case-based selection prior to the BACs removal stage are shown. The ROC curves show that the sensitivity of the CADe system with a case-based selection stage is higher than the CADe system over the whole specificity range. The area under the ROC curve between the two systems was significantly different, 0.91 ± 0.01 versus 0.89 ± 0.02 ($p<0.05$) over 5,000 bootstraps. Compared to the initial CADe system, the sensitivity of the CADe system without the case selection is higher only at a specificity of 0.92 or higher. At a lower specificity, the sensitivity of the CADe system without the case selection is lower than the initial CADe system. The area under the ROC curve between the systems was not significantly different, with an area under the curve of 0.90 ± 0.02 for the initial CADe system versus 0.89 ± 0.02 ($p = 0.30$) for the system without case-selection over 5000 bootstraps.

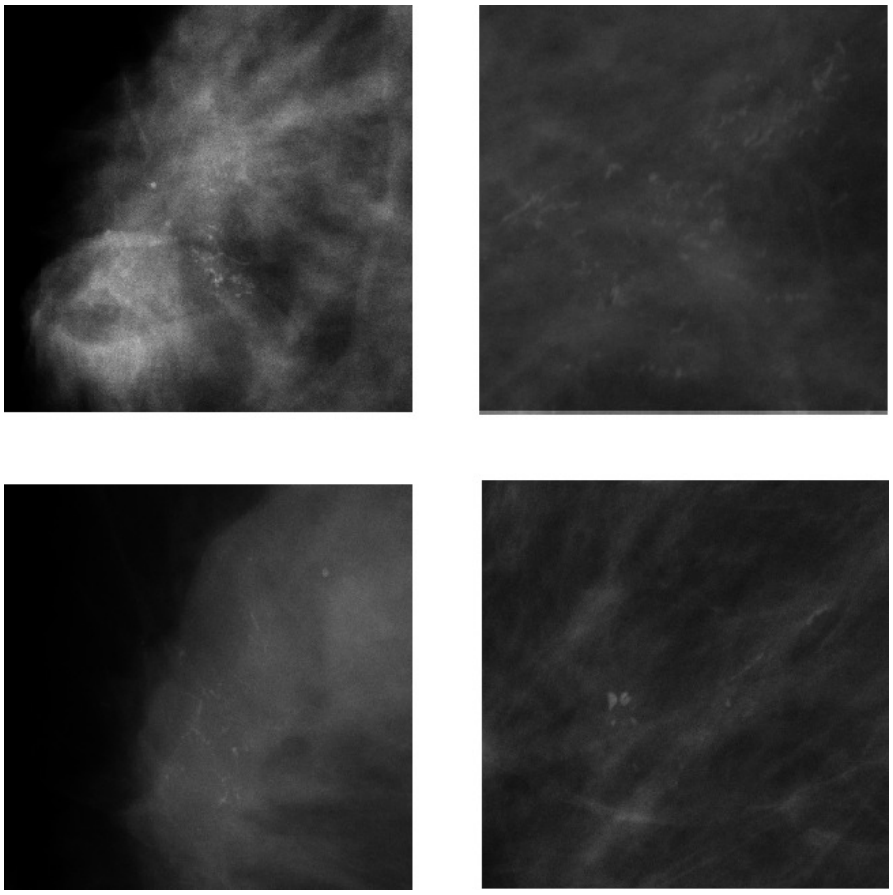


Figure 4.9: Examples of missed malignant calcification groups when the case-based selection was not performed prior to the BACs removal stage.

4.4 Discussion

In this study, we contributed to the development of a CADe system for the detection of calcifications. The ultimate aim of this work is to develop a standalone system which can compete with radiologists as an independent observer. Towards this goal, we proposed an additional stage to the CADe system where BACs are removed. Here, novel features are engineered to detect BACs and eliminate them as findings from the CADe system. The CADe system with BACs removal was evaluated on a test set which contained 3,574 mammographic images. These images came from 993 normal cases and 95 cases with malignant calcifications. The performance of the CADe system was compared to the CADe system without BACs removal. The results show a significant increase in sensitivity at operating points ranging from a specificity of 0.8 to 1.0 when BACs were removed.

At high false positive rates performance of CAD did not improve, which is understandable. The purpose of this study was to improve the CADe system at a high specificity (e.g. $>$0.9). Therefore, calcifications detected at a specificity of 0.75 or lower were not considered for the case-based selection and BACs removal stage. Results show that the sensitivity increases less at operating points close to this threshold value. This indicates that the false positives which are present at these relatively high false positive rates do not originate from BACs. An increase of the performance at higher false positive rates can be achieved when more false positives are taken into account, i.e. by lowering T_{Group}. However, the causes of the false positives in this spectrum are different and other types of calcifications will become predominant instead of BACs.

Also at very few false positives per case ($<$0.004 FP/C, one false positive per 250 cases), the sensitivity of the CADe system with BACs removal does not differ from the initial CADe system. This can indicate that the false positives with the highest CADe scores do not originate from BACs. However, there are only few false positives in our dataset determining the performance in this part of the FROC curve, which complicates accurate performance assessment of the CADe system.

In a previous study, the frequency of cases with BACs was assessed and it was shown that 52% of the cases with high CADe scores contained BACs[127]. In a previous study, the frequency of cases with BACs was assessed and it was shown that 52% of the normal cases with high CADe scores contained BACs[28]. In our validation set we found that 31% of the normal cases with CADe scores above T_{Group} contained BACs. This discrepancy might arise because in the previous study, T_{Group} was set to a higher value and a different dataset was used. The maximum decrease in FP/C between the CADe system with BACs removal and the initial CADe system was 44%.

This suggests BACs removal with our system is successful in the majority of cases. The small discrepancy may be caused by cases which contain only one (small) group of BACs. These cases are difficult to be selected by the CADe system. Furthermore, it is possible that in selected cases not all BACs clusters are removed in the BACs removal stage.

There is a possibility that malignant calcification groups can incidentally mimic BAC patterns. However, the BACs removal approach with case-based selection missed only one additional malignant calcification group compared to the initial CADe but this malignant calcification group was found by our system in the other view. When the case selection approach was not applied, the number of missed malignant calcification groups increased up to 23 groups causing misclassification errors of 2 additional malignant cases. These results suggest that, although malignant calcification groups can be linearly distributed, it is quite unlikely that malignant clusters are missed in both views by the removal process. The case-based selection process minimizes this risk even more. However, removing malignant calcification groups that might look like BACS can never be completely avoided. The same holds for radiologists who may also have difficulties with such cases. A thorough investigation of this issue would require a much larger database because malignant clusters that look similar to BACs are rare.

We found that applying a case-based selection before the removal stage is beneficial to prevent the removal of malignant findings. The system with the case-based selection before the BACs removal showed a significantly better performance than the system without a case selection. The number of cases which were analyzed by the BACs removal stage was determined by T_{Case}. This threshold can influence the performance of the system. For example, setting the threshold lower makes the selection less specific and more cases will go to the BACs removal stage. Additionally, T_{BACs}, the threshold for the BACs removal stage, also influences the system. Setting a low T_{BACs} might not make the BACs removal stage specific enough. After performing a grid search to optimize the two thresholds, T_{Case} and T_{BACs}, these thresholds were set to 0.5 and 0.5, respectively. However, it was found that, when taking non-extreme values for these thresholds, the end performance was very similar. Therefore, this procedure did not cause a bias in the results.

Because BACs are not considered as significant findings in mammography screening, only few studies have been carried out for the detection and classification of BACs[120,121,139]. These studies address the detectability and segmentation of BACs and do not report the effect on overall detection performance of CADe systems. Therefore, it is not possible to compare our results to these studies.

Although BACs are considered as irrelevant findings in breast cancer screening,

many studies have been done on the assessment of BACs and their relation to various pathologies. One of these pathologies is cardiovascular disease. However, there is still some controversy about the role of BACs in the development of this disease[14–25]. BACs have also been related to several other pathologies[26–30] and have even been related to breast cancer[140]. Therefore, automated detection of BACs in screening mammograms may be of interest outside the scope of CADe and might in the future be applied for detection and diagnosis of other pathologies than breast cancer.

In this paper, we showed a significant improvement in the detection of malignant calcifications. However, with a sensitivity of 76% at a specificity of 0.95, the performance of the CADe system does not yet seem comparable to the performance of screening radiologists. To compare the performance of a CADe system and a screening radiologist is not an easy task. In the first place, the recall rate based on suspicious calcifications has to be known as well as the positive predictive value of the recalled cases. In the second place, only exams which were recalled by the radiologists in screening were used in this study. For a fair comparison, exams prior to the recalled exams should be present in the dataset, because some false negatives of the radiologist might be found by the CADe system. Nevertheless, we still expect that the performance of the CADe is not comparable to screening radiologists because many false positives detected by the CADe system originate from benign calcifications. In this paper, we have focused on the removal of BACs while still 39% of the false positives are caused by other types of benign calcifications. Many of these are obvious benign calcifications which would not be classified as suspect by radiologists. Therefore, further research should be done in the detection and elimination of these types of calcifications. A proposed method is to classify each type individually as there might be too much variation between the types to be distinguished from malignant calcifications as one group.

4.5 Conclusion

Breast arterial calcifications are a major cause of false positives in CADe systems. In this paper, we have proposed a method to tackle this problem by including a case-based classification stage and a removal stage in a CADe system. We have shown that the removal of BACs as positive findings can reduce the number of false positives marked by the CADe system significantly.

Removal of obvious false positive calcification findings in mammograms

5

Jan-Jurre Mordang, Albert Gubern-Mérida, Alessandro Bria, Francesco Tortorella, Gerard den Heeten, and Nico Karssemeijer

Original title: Improving computer-aided detection assistance in breast cancer screening by removal of obviously false positive findings

Published in: Medical Physics, 2017

Abstract

Purpose: Computer-Aided Detection (CADe) systems for mammography screening still mark many false positives. This can cause that radiologists lose confidence in CADe, especially when many false positive are obviously not suspicious to them. In this study we focus on obvious false positives generated by calcification detection algorithms.

Methods: We aim at reducing the number of obvious false positive findings by adding an additional step in the detection method. In this step, a multi-class machine learning method is implemented in which dedicated classifiers learn to recognize the patterns of obvious false positive subtypes that occur most frequently. The method is compared to a conventional two-class approach, were all false positive subtypes are grouped together in one class, and to the baseline CADe system without the new false positive removal step. The methods are evaluated on an independent data set containing 1,542 screening examinations of which 80 exams contain malignant calcifications.

Results: Analysis showed that the multi-class approach yielded a significantly higher sensitivity compared to the other two methods ($p<0.0002$). At one obvious false positive per 100 images, the baseline CADe system detected 61% of the malignant exams while the systems with the two-class and multi-class false positive reduction step detected 73% and 83%, respectively.

Conclusions: Our study showed that by adding the proposed method to a CADe system the number of obvious false positives can decrease significantly ($p<0.0002$).

5.1 Introduction

Breast cancer screening programs, in which asymptomatic women are periodically invited for a mammographic exam, have been introduced in many countries to detect breast cancer at an early stage. Several studies have shown that early detection reduces breast cancer mortality in women over the age of 40 by 30%[34,35]. However, reading of screening mammograms is a tedious and difficult task, in which radiologists have to read large numbers of mammograms of which only a few contain cancers. Furthermore, early manifestations of breast cancer appearing as calcifications often have a very subtle appearance and their characteristics may be similar to patterns commonly seen in benign disease. Therefore, radiologists have to be very attentive and concentrated when reading mammograms. When their concentration decreases due to fatigue or distraction this may have potentially serious consequences, i.e. cancers may be missed[141].

To decrease the workload and to assist screening radiologists in reading mammograms, Computer-Aided Detection (CADe) systems have been developed. In the US, these CADe systems are already widely used in screening practice for over a decade[67]. A CADe system usually consist of two separate subsystems, one for detecting suspicious masses and one for calcifications. Abnormalities detected by CADe systems are marked in mammography workstations during reading sessions to avoid that lesions are overlooked.

Although several reader studies have shown that the detection rate of individual radiologists increases when a CADe system is used[73,94–96], there is no convincing evidence yet that the incorporation of CADe systems into the mammography reading workflow contributed to an overall improvement of screening performance in daily practice[98,101]. This disappointing result might be explained by the fact that CADe systems operate at a low specificity, because a high sensitivity is desired to ensure that lesions are not missed. Consequently, many false positives are marked by CADe systems. This can lead to (1) an increase in the number of women being unnecessarily referred for a clinical follow-up[73,96,97], (2) an increase in interpretation time of the mammograms[102], and (3) a loss in confidence in the CADe system[102], especially when locations are marked which are obviously not suspicious. CADe marks on regions that are evidently normal can easily be dismissed by radiologists. In this paper we will refer to them as obvious false positives (OFPs). Many OFPs are generated during the detection of calcifications. They can be categorized in three types: 1) macrocalcifications, 2) breast arterial calcifications (BACs), and 3) detection errors of the baseline system. Examples of the three OFP types, together with an example of a non-OFP, are shown in Figure 5.1.

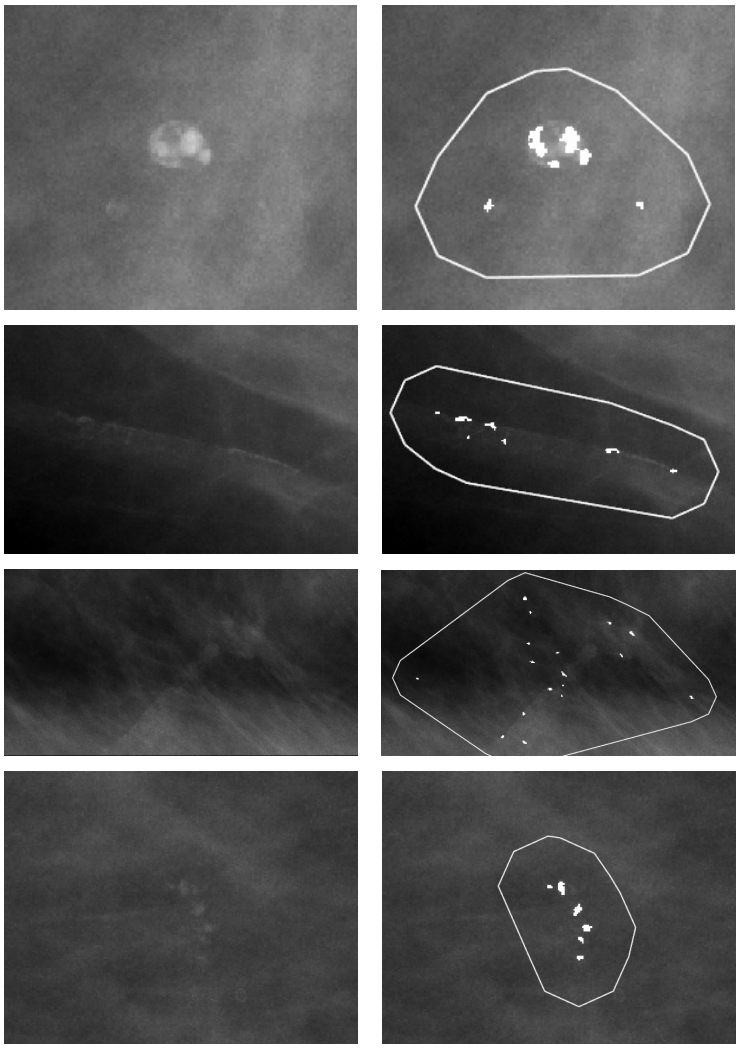


Figure 5.1: Examples of false positive findings. The first three examples are considered as obvious false positives, from top to bottom: macrocalcifications, breast arterial calcifications and detection errors. The bottom example is an example of a finding that is not an obvious false positive. Corresponding CADe findings are shown in the right column.

To understand why OFPs are detected one has to look at the design of current CADe systems. Most systems consist of three steps: 1) a calcification detection and segmentation step, 2) a clustering step, and 3) a classification step in which groups of calcifications are classified into malignant and non-malignant classes. The non-malignant class can contain benign calcifications as well as normal tissue. A single classifier is often used in the third step. For this purpose, various features are calculated to describe the two classes in the best possible way. While the samples included in the malignant class tend to have a homogeneous feature response, the non-malignant class includes different subtypes, such as the previously described OFPs. These have large differences in appearance causing a much more heteroge-

neous feature response within the non-malignant class. This may lead to a classifier which is not able to generalize well when applied to unseen test data.

In this study, we aim at improving the output of the CADe system by automatically reducing the number OFPs in an additional step. We propose two methods, 1) a method consisting of a multi-class approach in which dedicated classifiers learn the specific patterns of each OFP subtype and 2) a method with a conventional 2-class classification strategy where all OFPs are grouped as one class. For evaluation purposes, we incorporated the proposed method into a state-of-the-art calcification detection system and studied its contribution on an independent data set composed of 1,542 screening examinations.

5.2 Materials

Mammograms used in this work were collected in the Dutch Breast Cancer Screening Program (Bevolkings Onderzoek Midden-West, The Netherlands). All mammograms were acquired using Hologic digital mammography systems (Hologic, Bedford, Massachusetts, United States) and the "for processing" images were archived. This allowed us to work with the raw data. For training, we used a data set containing 1,837 screening exams from different women. In total, 6,119 Medio-Lateral Oblique and Cranial-Caudal views were included. Of the 1,837 exams, 1,670 exams did not contain any abnormalities and had at least 2 years of follow up available with no sign of breast cancer. These exams were considered as normal. The remaining 167 exams were recalled in screening and had a biopsy proven malignancy. These exams contained 336 groups of malignant calcifications which were all annotated based on the diagnostic reports.

In addition to the set described above, an independent test set of 1,542 screening exams was collected. Of these exams, 80 contained biopsy proven malignant calcifications and the remaining 1,462 screening exams did not contain any abnormalities. In the exams with malignancies, a total of 158 groups of malignant calcifications were annotated based on the diagnostic reports. This data set did not have any overlap with the training data set.

5.3 Methods

5.3.1 Framework

In this work, we propose an additional step within the CADe pipeline to improve the detection of calcifications by reducing the number of OFPs. This framework is

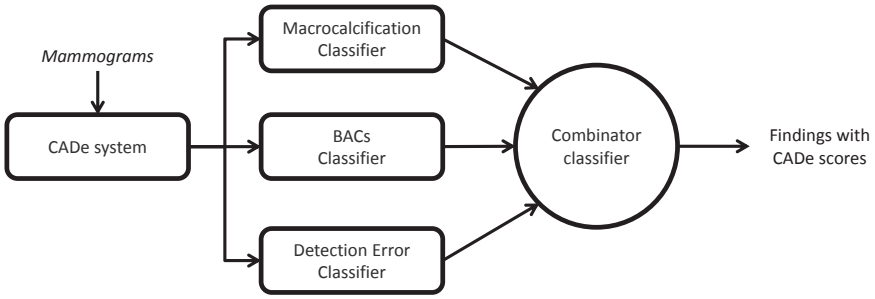


Figure 5.2: The proposed framework: The input mammogram is processed by the CADe system based on the algorithm of Bria et al[74]. Subsequently, three different scores representing the likelihood of each finding being a macrocalcification, BAC or a detection error are computed using three dedicated classifiers, respectively. The final suspiciousness score of the multi-class method is computed by combining the output of the three dedicated OFP systems.

shown in Figure 5.2. First, CADe findings are computed on the input mammogram using a previously developed CADe system[74] which we will refer to as the baseline method. CADe findings are groups of calcifications with a suspiciousness score. Next to these scores, findings also contain the location of each detected calcification within the group and likelihood scores representing the confidence of the baseline system that these truly are calcifications. Examples of CADe findings are shown in Figure 5.3.

In the proposed novel step of the enhanced CADe system, for each CADe finding three additional scores are computed. These respectively represent the likelihood that a finding is caused by a macrocalcification, by breast arterial calcifications (BACs), or by a detection error of the baseline system due to noise or normal tissue patterns. These three scores are calculated using three dedicated classifiers. A final suspiciousness score of this multi-class approach is recomputed by combing the output of the three dedicated OFP classifiers.

5.3.2 CADe system

The baseline CADe system used in this work is based on the method described in Bria et al[74] and consists of three steps. First, calcifications are detected on the input mammogram. Here, a probability map is generated using a cascade classifier. This classifier consists of a series of GentleBoost[142] classifiers trained on Haar-like features[79,80]. Thresholding and connected component analysis is performed to obtain segmented calcification candidates. In the second step, groups of calcifications are

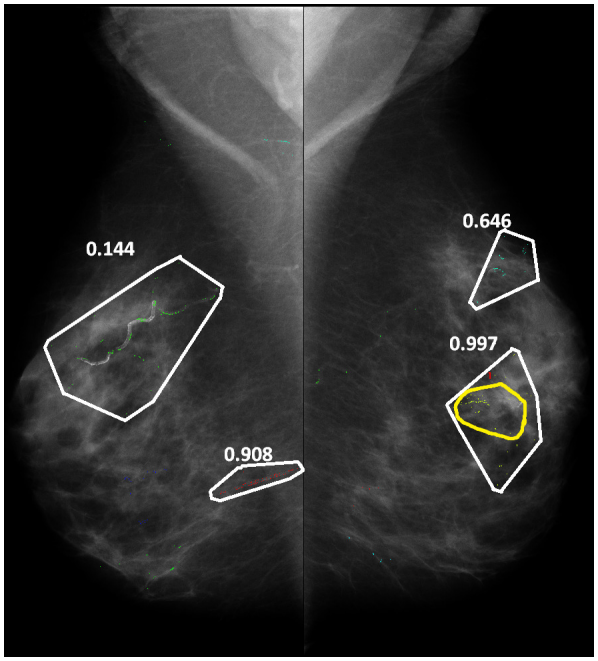


Figure 5.3: Example of CADe findings. In the image, CADe findings are shown together with their suspiciousness scores (white annotations). Additionally, an annotation of malignant calcifications (yellow annotation) is shown.

created by clustering candidates that lie within 10mm from each other. Finally, a single GentleBoost classifier is employed to differentiate malignant calcification groups from benign groups resulting in a suspiciousness score for each group. In our study, the baseline CADe system was trained on the same data set described in Bria et al[74].

5.3.3 OFP removal step

After applying the baseline CADe system, resulting CADe findings are processed independently by three dedicated classifiers to compute the likelihood of being an OFP subtype. These three dedicated classifiers are trained to discriminate malignant findings and OFP subtypes. A set of features to describe CADe findings is used. These features can be categorized in three levels: i) calcification features, ii) calcification group features, and iii) exam features. The set of features used in the study was designed to describe each finding in terms of their shape, topology, probability of the individual calcification detection step, texture and vesselness[74,81,143].

The smallest scale for the calculation of features is the calcification level. Feature values are calculated for each individual calcification within the finding. Then, feature values for the whole finding are computed as the mean, maximum, minimum and standard deviation of the calcification feature values. The calcification features are listed in Table 5.1. These features are designed to capture the wide variation in

shape and intensity among calcifications and their surrounding.

The second level is at the scale of the finding itself, i.e. a group of calcifications. Features used to characterize the groups are listed in Table 5.2. The main purpose of these features is to capture the spatial distribution of the calcifications and the underlying texture of the breast.

The highest level we consider for definition of features is the exam. Here information of all findings in the exam is combined. The main purpose of these features is to describe frequency and distribution patterns of findings in both breasts. The exam features are listed in Table 5.3.

In the final step of the OFP removal approach, the output of the three classifiers is combined to generate a suspiciousness score for each finding. Here, a single classifier was trained using the output of the three dedicated classifiers as features to discriminate between malignant and OFP findings.

5.4 Evaluation and Experiments

5.4.1 Evaluation

For evaluation of the experiments, Receiver Operating Characteristics (ROC) and Free-response ROC (FROC) analysis were performed at region and exam levels. At the region level, a finding detected by the system was considered a true positive when at least 2 of the detected calcifications within the finding where located in the ground truth annotation. There was no linking between regions that could be seen in multiple views and were considered as different regions. When multiple CADe findings hit the same annotation, the finding with the highest CADe score was chosen and the rest was ignored. Ground truth annotations that did not coincide with any CADe finding were considered as false negatives. Obvious false positives were computed as the number of CADe findings detected in normal exams which were labeled as an OFP finding. At the exam level, a true positive exam was an exam with at least one CADe finding hitting an annotation. If multiple annotations were hit, the exam-based score was set to the highest scoring finding. Normal exams were considered as false positives when at least one OFP finding was detected. Similarly, the highest scoring OFP finding was set as the exam-based score.

Statistical comparison was performed by means of bootstrapping[90]. Here, the test set was bootstrapped 5000 times and ROC and FROC analysis was carried out for each bootstrap.[113,114] Each bootstrap was constructed by sampling the data set with replacement. From the ROC and FROC curves, areas under the curve (AUCs) and partial areas under the curve (pAUCs) were calculated, respectively. For test-

Calcification features	
Feature Type	**Description**
Perimeter	The number of pixels that touch a background pixel with at least one side.
Area	The number of pixels belonging to a detected calcification.
Compactness	$p^2/(4 * \pi * a)$, p = perimeter, a = area.
Eccentricity	$\frac{I_{xx}+I_{yy}-\sqrt{(I_{xx}-I_{yy})^2+4I_{xy}^2}}{I_{xx}+I_{yy}+\sqrt{(I_{xx}-I_{yy})^2+4I_{xy}^2}}$ where I_{xx}, I_{yy} and I_{xy} are the moments of inertia.
Thickness	Width of the best fitting rectangle.
Ellipse	Ratio between the short axis and long axis of the best fitting ellipse.
Direction	Direction in which the calcification is located viewed from its clusters gravity center with respect to the x-axis.
Distance to centroid	Distance to the center of the finding.
Distance to nearest	Distance to the nearest calcification.
Distance to skin-air	Distance to the skin-air boundary.[81]
Mean edge	Mean output of a canny edge detector.
Degree	Number of edges incident to the calcification.
Normalized degree	Sum of the normalized weights of the edges incident to the calcification.
Probability (4)	Maximum, second maximum, mean, and standard deviation of the probability map of the calcification detection step.
Background (2)	Mean and standard deviation of the background pixel values, where the background is defined as pixel thick band surrounding the calcification.
Foreground (2)	Mean and standard deviation of the foreground pixel values, i.e. pixels within the calcification.
Contrast (5)	Maximum, mean, standard deviation, kurtosis, and skewness of $C_i = Log(y_i) - Log(y_b)$, where y_i is calcification pixel intensity of pixel i and y_b is the mean background intensity.
Attenuation	Maximum, mean, standard deviation, kurtosis, and skewness of $\Delta\mu = \mu_b - \mu_{mc} = C_i/d_i$, where d_i is the thickness of the calcification at location i
Image moments	Seven Hu invariants[144]

Table 5.1: Features based on the segmented calcifications in a finding. For each finding, the mean, max, min and standard deviation of the distribution of calcification feature values are used as features for the classifier.

Finding features	
Feature Type	**Description**
Area	The area of the finding computed on the convex hull[136] fitted on the whole finding.
Eccentricity	Eccentricity on the convex hull of the finding.
Ellipse	The ratio between the long axis and the short axis of a ellipse fitted on the convex hull of the finding.
Calcification number	The number of calcifications in the finding.
Density	$\frac{2\|E\|}{n(n-1)}$ where E is the number of edges of the graph made from the centers of each calcifications within the finding.
Coverage	$\sum_{i=1}^{n} \frac{A_{mC_i}}{A_{cls}}$ where A_{mC_i} is the area of the calcification i, n the number of calcifications within A_{cls}, the cluster area.
Orientation	The orientation of the cluster with respect to the xy-plane.
Circle coverage	The ratio of the surface area of all calcifications and the area of the minimum enclosing circle
Elongatedness	The ratio of the eigenvalues calculated with principle component analysis on subgroups with 10mm radius
Distance to skin/air	The distance of the center of the cluster to the skin air boundary.
Probability	Finding probability from the cluster detection (intermediate score of the CADe system).
2nd step score	The score given by the CADe system
Hessian (5)	The determinant of the Hessian matrix for each pixel.
Tubeness (5)	$k_{line}(\lambda_1, \lambda_2) = \frac{\lambda_2 - \lambda_1}{\lambda_2}$ where $\lambda_1 \leq \lambda_2$, the absolute eigenvalues calculated.
Lambda (5)	The highest absolute eigenvalue λ_2.
Vesselness (5)	Vesselness image filtering technique proposed by Frangi et al[145].

Table 5.2: Features representing the spatial organization of findings. The Hessian, Tubeness, Lambda, and Vesselness features are calculated at several scales: $0.2 \leq \sigma \leq 1.0$, steps of 0.2 resulting in 5 feature values per feature type.

Exam features	
Feature Type	**Description**
Findings	The number of findings in the exam.
FindingsHigh	The number of findings in the exam with a suspiciousness score >T.
FindingsCurrentSide	The number of findings at the breast side of the finding (left or right breast)
FindingsCurrentSideHigh	The number of findings at the breast side of the finding where only findings with a suspiciousness score >T are considered
FindingsContralateralSide	The number of findings at the breast side other than the location of the finding (left or right breast)
FindingsContralateralSideHigh	The number of findings at the breast side other than the location of the finding where only findings with a suspiciousness score >T are considered

Table 5.3: Features based on all findings within the exam.

ing significant differences, one-way ANOVA was applied on all bootstraps and a Bonferroni correction was performed to correct for multiple comparisons. Statistical analysis was carried out in SPSS (Version 22.0. IBM Corp. Armonk, New York, United States). Differences were considered to be significant for p-values <0.05.

5.4.2 Visual assessment of false positive CADe findings

In the first experiment, we visually inspected the different types of false positives generated by the baseline CADe system described in Section 5.3.2. For this purpose, this system, which was trained on an independent data set, was applied to the study data set. Exam-based FROC analysis was performed and a subset of the detected false positive findings was selected by choosing an operating point on the FROC curve. This operating point was chosen as such that the size of the selected set was manageable for visual inspection while the sensitivity was high enough to obtain a representative sample. The exam-based sensitivity we used was 0.91, which is similar to the sensitivity of commercially available CADe systems[74].

Each of the selected false positive findings was labeled as one of the following categories: 1) macrocalcifications: a finding containing detected calcifications which are (part of) macrocalcifications (i.e. calcifications larger than 1mm); 2) breast arterial calcifications (BACs): a finding containing detected calcifications located in an artery; 3) detection errors: a finding that does not contain any calcifications; and 4) other benign calcifications: groups of calcifications which contain benign calcifications other than macrocalcifications and BACs. Labeling of the false positives was performed by a researcher with experience in reading mammograms. In this work we focus on removing obvious false positives. Therefore, the fourth group of false positives was not used and ignored in the rest of the experiments.

5.4.3 Obvious false positive classification

In the second experiment, we investigated the classification performance of the two OFP removal approaches in discriminating malignant findings from OFPs. Based on the visual assessment performed in the previous experiment, we created a data set of CADe findings that mark malignant calcifications and OFPs.

For classification of malignant regions and OFPs, three GentleBoost classifiers, using regression stumps[78] as weak classifiers (n=100), were trained as dedicated OFP classifiers, one for each different OFP subtype. Each individual OFPs classifier was trained using all features described in Section 5.3.3. The threshold for the exam features was set to the same operating point as the visual assessment experiment. The combinator classifier was trained on the three (raw) output scores of the

dedicated classifiers, which values range from 0.0 to 1.0. For training the combinator classifier, several classifier types were used to investigate impact of the choice of this classifier on the final result: a GentleBoost classifier, a linear discriminant classifier, a Support Vector Machine[146] (SVM), and a Random Forest[147]classifier. For the GentleBoost classifier we used 100 regression stumps and for training of the Random Forest the maximum number of trees was set to 100 and the maximum tree depth was taken as the square root of the number of samples. For the SVM classifiers, three different kernels were assessed: a linear kernel, a polynomial kernel, and a radial basis function (RBF). The parameters for the SVM classifier with a polynomial and RBF were estimated by splitting up the training data into a training set and validation set[148]. Estimation was done with a grid search over C values ($C = 2^{-5}$, 2^{-3},...,2^{9}) and γ values ($\gamma = 2^{-15}$,2^{-13},...,2^{0}). The degree of the polynomial kernel was set to 3. The classifier for the conventional two-class classification of OFPs and malignant findings consisted of a GentleBoost classifier with 100 regression stumps.

The evaluation in this experiment consists of three parts, 1) the performance of the two approaches in discriminating each individual type of OFP from malignant findings, 2) the influence of the various combinator classifiers on the classification of OFPs and malignant, and 3) classification performance comparison between the novel multi-class approach, the conventional two-class method, and the baseline CADe system.

For evaluation of the multi-class approach and the two-class classification method, ROC analysis was performed with 10-fold cross-validation. Folds were made on case-level such that there was no overlap between training and testing, i.e. images of a given woman were either in the test or the training set but never in both. Furthermore, we ensured that all malignant and OFPs findings were equally distributed across all folds. For each fold, 9 folds were used for training of the classifiers and the remaining fold was used for testing the classifier. Region-based ROC analysis was applied on the output scores and AUC values were calculated.

5.4.4 CADe with obvious false positive removal

In the final experiment, we investigated the performance of the CADe system on detecting exams with malignant calcifications when adding the OFP removal step. First, the baseline system was applied to the independent test set. Then all findings above the previously defined operating point were processed by the proposed OFP removal step. On the resulting finding scores, exam-based ROC and FROC analysis was performed. For the ROC and FROC analysis, benign calcifications that were not categorized as OFP were ignored.

False positive distribution	
False positive type	**# findings (% of total)**
Macrocalcifications	82 (6%)
BACs	429(29%)
Detection errors	189(13%)
Other benign calcifications	756 (52%)

Table 5.4: Distribution of false positives (n = 1456) into 4 groups after visual analysis

On the test set, pAUCs were calculated for statistical evaluation because in this experiment only findings above the specified threshold were analyzed. The pAUCs where calculated on 5000 bootstraps and were pairwise compared to each other. Bonferroni correction was applied to correct for three comparisons: the baseline CADe system versus the two-class method, the baseline CADe system versus the multi-class method, and the two-class method versus the multi-class method. For the ROC analysis, the pAUC was calculated in a specificity range of 0.77 to 1.00 to compare the three methods. The exam-based FROC analysis was carried out to determine the sensitivity in terms of the number of obvious false positives per image. For comparison, the pAUC was calculated between 0.001 obvious false positives per image (one obvious false positive per 1000 images) and 0.02 obvious false positives per image (one obvious false positive per 50 images) for the three methods. We chose these obvious false positive per image operating points because we are interested in high specificity range and these points represent the same specificity range of radiologists in screening.

5.5 Results

5.5.1 Visual assessment of false positive CADe findings

In total, 1,456 of the 20,390 (7%) false positive findings and 265 of the 336 (76%) malignant findings had a score above the threshold, while 700 (48%) of the false positives were labeled as OFPs. The distribution of the false positives over the previously defined subtypes is shown in Table 5.4. After excluding the benign calcifications, 965 findings remained. These were used in the experiments below.

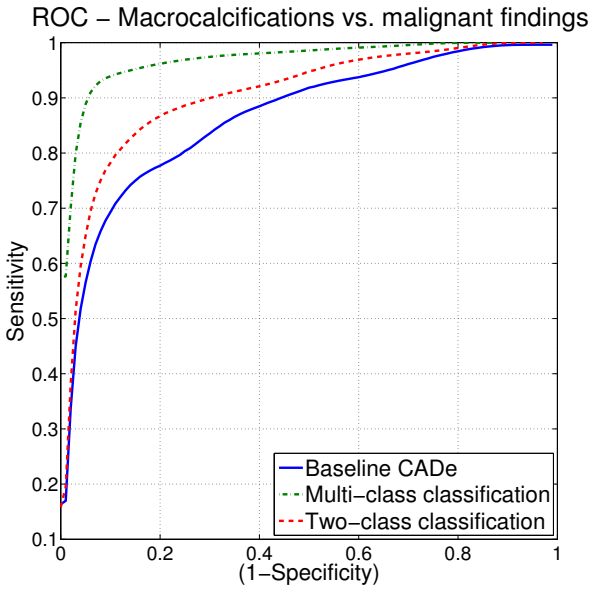


Figure 5.4: ROC curves for macrocalcification versus malignant findings. The average curves are shown computed from 5000 bootstraps.

5.5.2 Obvious false positive classification

Performance on individual OFP subtype classification

For each OFP type, an individual classifier was trained. These dedicated classifiers were compared to the output of the baseline CADe system and to the conventional two-class classification method. The ROC curves for each OFP type are shown in Figures 5.4-5.6. In each subfigure, three curves are plotted for the three methods. These curves show that the sensitivity is higher in the whole specificity range for both the two-class classification and the multi-class classification method compared to the baseline CADe system. Additionally, the multi-class classification results in a higher sensitivity compared to the two-class classification. The corresponding AUC values of each ROC curve are shown in Table 5.5. Here, it can also be observed that the two-class and multi-class classification methods result in significant higher AUC values than the baseline CADe system ($p<0.0002$).

Performance of the combinator classifier

Several classifiers were trained to combine the three output scores of the individual OFPs classifiers. We compared the AUC values from the ROC analysis on the whole OFP finding set and malignant findings between the GentleBoost, Random Forest, and SVM classifiers. The AUC value for each classifier type is shown in Table 5.6. The differences in AUC values of these classifiers were not large (AUC values range from 0.880 to 0.937). In the end, we choose the linear SVM in the next experiments

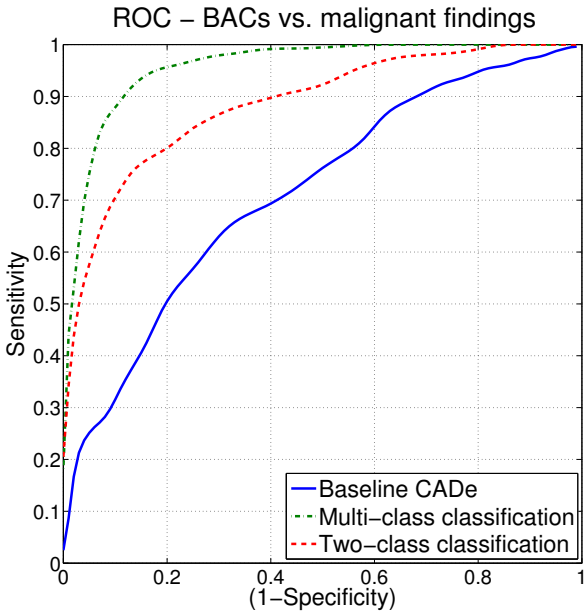


Figure 5.5: ROC curves for BACs versus malignant findings. The average curves are shown computed from 5000 bootstraps.

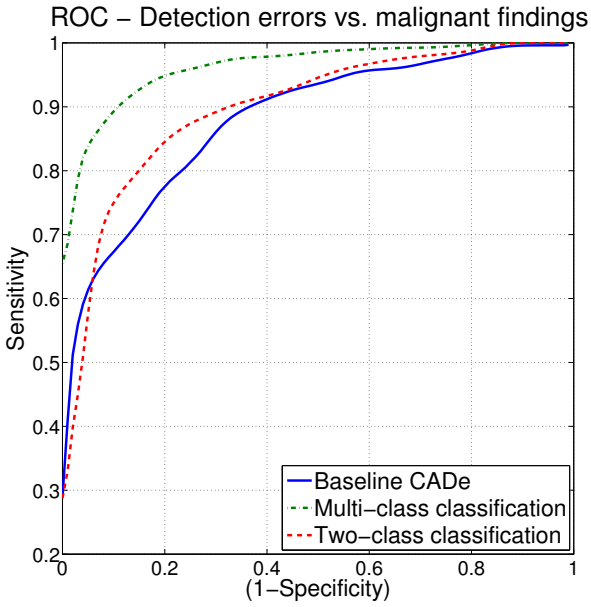


Figure 5.6: ROC curves for detection errors versus malignant findings. The average curves are shown computed from 5000 bootstraps.

AUC values of individual OFP classification			
Method	**Macrocalcifications**	**BACs**	**Detection errors**
Baseline CADe	0.866 ± 0.015	0.715 ± 0.022	0.880 ± 0.017
Two-class classification	$0.916 \pm 0.018^{\dagger}$	$0.936 \pm 0.010^{\dagger}$	$0.910 \pm 0.015^{\dagger}$
Multi-class classification	$0.970 \pm 0.010^{\dagger\ddagger}$	$0.957 \pm 0.007^{\dagger\ddagger}$	$0.963 \pm 0.009^{\dagger\ddagger}$

Table 5.5: Area under the ROC curve for discrimination between malignant findings and macrocalcifications, BACs and detection errors. Average AUCs (mean±stdev) computed from 5000 bootstraps. All p-values in all comparisons where <0.0002. † $p<0.05$ when comparing two-class and multi-class classification to the baseline CADe score.‡ $p<0.05$ multi-class classification compared to the two-class classification

AUC values of OFPs classification various combinator classifiers					
GentleBoost	**LDA**	**Random Forest**	**SVM**		
			Linear	*Polynomial*	*RBF*
0.924	0.930	0.880	0.937	0.927	0.928

Table 5.6: Area under the ROC curve for the classification of findings into malignant or OFPs (macrocalcifications, BACs and detection errors) for all four classifier types. Average AUCs are shown of 5000 bootstraps.

since this classifier obtained the best performance.

Classification performance comparison OFP versus malignant

In Figure 5.7, the ROC curves of the classification of all OFPs versus malignant findings are shown for the three methods. The ROC curves for the two-class and multi-class classification methods resulted in a higher sensitivity than the baseline CADe system over the full specificity range. The corresponding AUC values are shown in Table 5.7. The AUC values of the two-class and multi-class classification methods are significantly different from the AUC value of the baseline CADe system: 0.777 ± 0.018 versus 0.927 ± 0.010 and 0.937 ± 0.010 ($p<0.0002$ in both comparisons), respectively. The AUC value for the multi-class classification method is higher than the two-class method but the difference was not significant ($p = 0.07$).

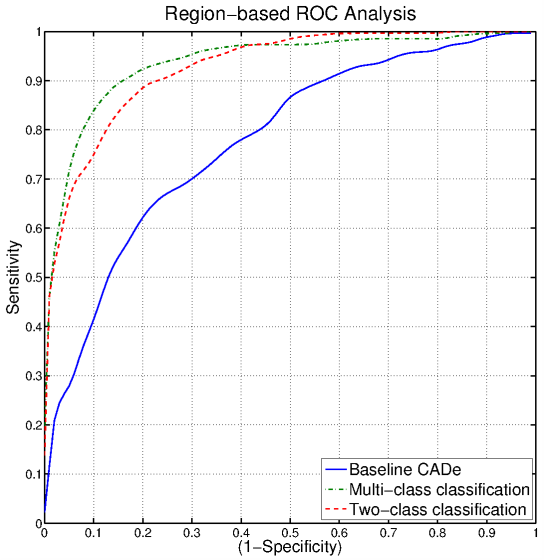


Figure 5.7: ROC curves for the classification of all OFPs versus malignant regions for the three methods. Average curves are shown of 5000 bootstraps.

AUC values of OFPs classification with 3 methods		
Method	**AUC values**	**p-values**
Baseline CADe	0.777 ± 0.018	
Two-class classification	0.927 ± 0.010	$<0.0002^{\dagger}$
Multi-class classification	0.937 ± 0.010	$<0.0002^{\dagger}$ and 0.07

Table 5.7: Area under the ROC curve the classification of malignant findings and all OFPs (macrocalcifications, BACs and detection errors) for all three methods. Average AUCs are shown of 5000 bootstraps. † $p<0.05$ when comparing two-class and multi-class classification to the baseline CADe score.

5.5.3 CADe with obvious false positive removal

The performance comparison for the whole data set is shown in Figure 5.8. Because the threshold was set at a specificity of 0.77, the curves are equal at lower specificity. Therefore, the sensitivity is plotted for all three methods in a specificity range between 0.77 and 1.0. The Figure shows that the curves for the two-class and multi-class classification methods have a higher sensitivity over a specificity range of 0.8 to 1 compared to baseline CADe system. The pAUC in the specificity range of 0.77 to 1.0 are shown in Table 5.8. The two-class and multi-class classifiers have a significantly higher pAUC compared to the baseline CADe system ($p<0.0002$ in both comparisons). The multi-class classification resulted also in a significantly higher pAUC compared to the two-class classification ($p<0.0002$).

The exam-based FROC curves are shown in Figure 5.9. It can be seen that the two proposed methods have a higher sensitivity at 0.12 obvious false positive per image ($\approx$1 obvious false positive per 8 images) or less. At one obvious false positive per 100 images, the baseline CADe system has an exam based sensitivity of 61% in malignant exams, while the other two methods have exam-based sensitivities of 73% (two-class classification) and 83% (multi-class classification), respectively. At a sensitivity of 80%, the number of obvious false positives detected by the baseline CADe system is 0.038 per image ($\approx$1 per 26 images) while the two-class classification detects 0.0135 obvious false positives per image ($\approx$1 per 74 images) and the multi-class classification 0.0085 obvious false positives per image ($\approx$1 per 118 images). The pAUC values calculated over the range of 0.001 to 0.2 obvious false positives per image are shown in Table 5.9. These values are significantly different between the 3 methods. Both the two-class and multi-class classification methods have a significantly higher pAUC than the one obtained by the baseline CADe system ($p<0.0002$ in both comparisons). The pAUC of multi-class classification is significantly higher than the two-class classification ($p<0.0002$).

5.6 Discussion

In this paper, we have constructed a method for the removal of false positives which are obviously not suspicious when observed by a screening radiologist. Two methods are proposed to remove these obvious false positives. The first method is based on a single classifier trained specifically to discriminate between OFP and malignant findings. The second method consists of three classifiers to individually classify each OFP subtype and a fourth classifier to combine their outputs to discriminate between OFP and malignant findings. These two methods were compared to the baseline

pAUC values of the exam-based ROC with 3 methods		
Method	**pAUC values**	**p-values**
Baseline CADe	0.183 ± 0.008	
Two-class classification	0.199 ± 0.007	$<0.0002^{\dagger}$
Multi-class classification	0.205 ± 0.007	$<0.0002^{\dagger}$ and $<0.0002^{\ddagger}$

Table 5.8: Area under the exam-based ROC curve the classification of exams with malignant findings in the range of a specificity of 0.77 to 1.0 . Average AUCs are shown of 5000 bootstraps. † $p<0.05$ when comparing two-class and multi-class classification to the baseline CADe score. ‡ $p<0.05$ multi-class classification compared to the two-class classification.

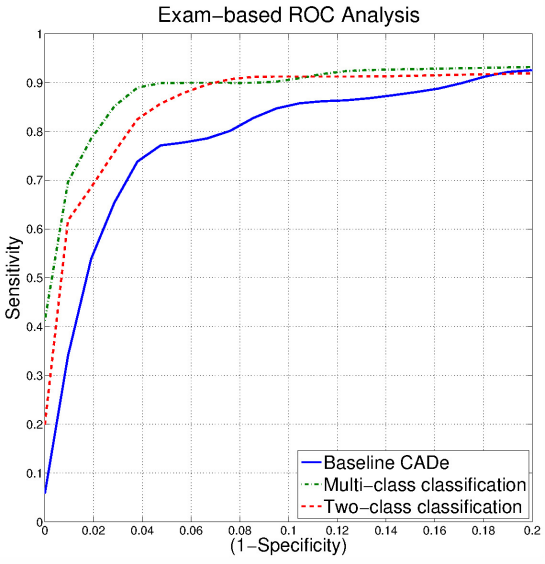


Figure 5.8: Exam-based ROC curves for the classification of the whole data set for the three methods in a specificity range of 0.8 - 1.0. Average curves are shown of 5000 bootstraps.

pAUC values of the exam-based FROC with 3 methods		
Method	**pAUC values**	**p-values**
Baseline CADe	0.220 ± 0.016	
Two-class classification	0.256 ± 0.015	$<0.0002^{\dagger}$
Multi-class classification	0.268 ± 0.012	$<0.0002^{\dagger}$ and $<0.0002^{\ddagger}$

Table 5.9: Area under the exam-based FROC curve the classification of exams with malignant findings in the range of 0.001 to 0.2 obvious false positives per image. Average AUCs are shown of 5000 bootstraps.† $p<0.05$ compared to the baseline CADe score. ‡ $p<0.05$ compared to the classifier trained on all OFPs.

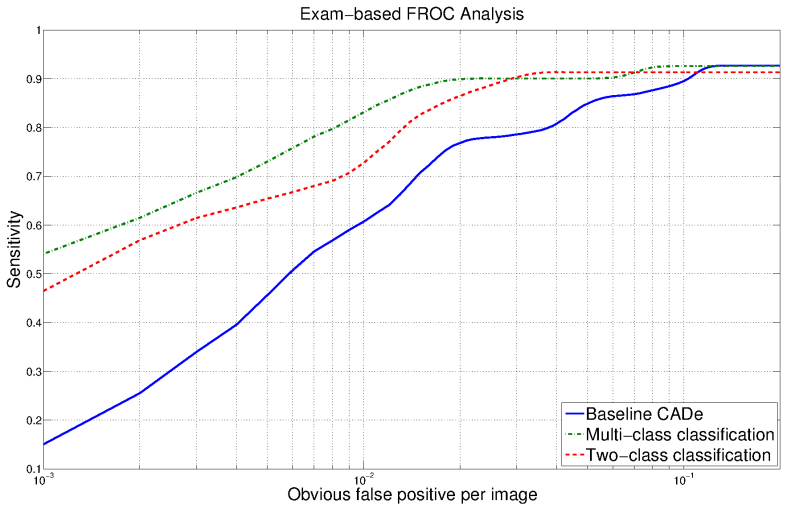


Figure 5.9: Exam-based FROC curves for the classification of the whole data set for the three methods in an obvious false positive per image range of 0.001 - 0.2. Average curves are shown of 5000 bootstraps.

CADe system using a large test set containing 80 exams with 158 annotated malignant calcification groups and 1,462 exams without any malignancies. The results show that the additional step helps improving the overall CADe performance.

In this study, we have only focused on the obvious false positive findings, i.e. findings that would be easily dismissed as being benign by a screening radiologist, because these are the false positive findings that cause loss of confidence in CADe systems. The visual assessment performed in this study showed that approximately half of the false positives generated by the baseline CADe system belong to this class (48%). The other false positives were categorized as other benign calcifications. We ignored this class because, when using CADe in combination with radiologists, this type of benign findings can be considered as less disturbing and it might in fact be debatable whether these benign findings should considered as true or false positives. Note that, many of these benign calcifications might be suspicious and in practice require further investigation. However, since we develop a standalone CAD system with high sensitivity for detection of breast cancer with as few false positives as possible, it can be argued that in our application the detection of these benign calcifications should not be rewarded.

Classifying each OFP subtype with a separate classifier trained on that specific type resulted in a higher sensitivity than training one classifier with all OFPs grouped together. This can be explained by the fact that there is much variation between the features computed for each OFP subtype. Such a large variation between feature responses is challenging for a single classifier, which might have problems generalizing among the OFP subtypes, resulting in a suboptimal classification. On a finding scale, combining the output of the three classifiers led to a higher AUC when compared to a single classifier discriminating between OFPs and malignant findings. This difference was not found to be significant ($p = 0.07$). Both the multi-class and two-class approaches significantly improved the classification compared to the baseline CADe system ($p<0.0002$).

Various classifiers have been evaluated in this study to differentiate the multi-classes such that malignant regions and OFPs are separated in the most optimal way. However, further research can be performed in optimizing the final combinator classifier by assessing other classifiers or by applying a likelihood score scaling transformation to the output of the individual classifiers to normalize the scores for training the combinator classifier[149,150].

We found that adding a classification step to the CADe system for the removal of OFPs leads to a significant increase in overall exam-based sensitivity, in which the CADe finding with the highest score was taken for the whole exam ($p<0.0002$). At a specificity of 0.95, the sensitivity for detecting exams with malignant findings in-

creased from 77% up to 86% when a classifier which is trained on all obvious false positives is added. The improvement was even higher (up to 90%) when we applied the multi-class approach. A larger difference in sensitivity was observed at higher specificity. The maximum difference in sensitivity can be observed at a specificity of 0.99 where the baseline CADe system detected 34% of the malignant exams and the two-class and multi-class classification schemes detected 62% and 70% of the malignant exams, respectively. Since a threshold was set at a specificity of 0.77, no differences can be observed at this operating point or at a lower specificity values.

There are similar studies presenting other classification approaches to discriminate between benign findings and detected malignant calcifications. In Veldkamp et al. (2000)[151], a k-nearest-neighbor classifier was used which resulted in an AUC of 0.83. In this study, however, obvious false positives were excluded. Wei et al. (2005)[152] evaluated several classifiers: a SVM, a kernel Fisher discriminant, and a relevance vector machine. The SVM method was the best performing classifier with an AUC of 0.85. In Elan et al. (2014)[153] a Circular Complex-valued Extreme Learning Machine was used for classification and a AUC of 0.96 was reported. However, in these studies the number of benign calcification samples was rather low and the origin of the samples in the benign class were not assessed or reported. Furthermore, the authors did not evaluate the performance of their algorithms when applied to a population representative of screening. It should be noted that, in our study, we have tried to mimic a screening situation by adding a large set of normal exams in the data set (we used a ratio of ten normals to one malignant exam).

In this study we used the algorithm presented in Bria et al. as our baseline CADe system. In their work, the authors showed that this algorithm performed at least as good as a widely used commercially available system on a comparable series of cases. We showed that its performance can be significantly improved by adding the proposed OFP classification step ($p<0.0002$). We believe that a similar strategy might be suitable to improve the performance of similar CADe systems.

CADe systems currently used in clinical practice operate at around 1 false positive per exam (assuming 4 mammographic images per exam)[74]. The baseline system system used in this study obtained a sensitivity of 91% at this false-positive rate. In the FROC analysis, it can be seen that the proposed approach considerably decreases the number of obvious false positives with a minor sensitivity loss. We believe that this reduction in false positives is important to improve radiologists' confidence in CADe systems. Furthermore, it should be noted that the ultimate goal of our research is to create a CADe system that would perform equal or better than a screening radiologist and can be used as an independent observer. In this scenario, assuming that we aim at a recall rate of 5%[125] of this system, the CADe system would

need to operate at a specificity of approximately 98% for calcification detection, because other algorithms for detection of masses will generate false positives as well. At this operating point, the baseline system achieves a sensitivity value of 54%. The proposed systems significantly improve the baseline method and obtain a sensitivity of 68% and 78% with the two-class and multi-class methods, respectively ($p<0.0002$ for both methods). However, this study was limited to the OFPs and the improvement in performance might not be as large when other benign false positives are also considered.

One of the limitations of our study is that we used screening data acquired on mammographic units of only one manufacturer. Mammograms of different manufacturers have different characteristics. Therefore, it is easier to make a CAD system work well on images from a single manufacturer than in a multi-vendor environment. This issue will be addressed in future research.

The performance of the system is not yet comparable to that of a single radiologist. Our proposed system detects 78% of the exams with malignant calcifications at a specificity of 98% and this sensitivity will decrease when including all benign findings in the validation. However, it should be noted that the proposed framework was validated with malignant recalled exams from screening judged by two radiologists, because double reading is practiced in the screening program. Comparison of our method with a single radiologist is not possible with the data we have, because we do not have access to the reports of the individual readers. From the literature it is known that double reading improves sensitivity by 13%[55]. This would mean that the sensitivity of a single radiologist would be 87% on our dataset, which still compares favorably to the performance of our CAD system.

5.7 Conclusion

In this paper, we have introduced a method to suppress false positive findings which are obviously not malignant to increase the benefit of using CADe systems in breast cancer screening. By adding our method to an existing CADe system, the number of false positives due to the obvious false positives strongly decreased leading to less false positives at a high sensitivity. The best results were obtained when applying a multi-class method with dedicated classifiers that learn the characteristic of each OFP subtype independently.

Part II

Evaluation of CAD and breast cancer screening

Assessment of the screening sensitivity for detection of malignant calcifications

6

Jan-Jurre Mordang, Albert Gubern-Mérida, Alessandro Bria, Francesco Tortorella, Ritse Mann, Mireille Broeders, Gerard den Heeten, and Nico Karssemeijer

Original title: The importance of early detection of calcifications associated with breast cancer in screening

Published in: Breast Cancer Research and Treatment, 2017

Abstract

Purpose: The aim of this study was to assess how often women with undetected calcifications in prior screening mammograms are subsequently diagnosed with invasive cancer.

Methods: From a screening cohort of 63,895 women, exams were collected from 59,690 women without any abnormalities, 744 women with a screen-detected cancer and a prior negative exam, 781 women with a false positive exam based on calcifications, and 413 women with an interval cancer. A radiologist identified cancer-related calcifications, selected by a computer-aided detection system, on mammograms taken prior to screen-detected or interval cancer diagnoses. Using this ground truth and the pathology reports, the sensitivity for calcification detection and the proportion of lesions with visible calcifications that developed into invasive cancer were determined.

Results: The screening sensitivity for calcifications was 45.5%, at a specificity of 99.5%. A total of 68.4% (n=177) of cancer-related calcifications that could have been detected earlier were associated with invasive when diagnosed.

Conclusions: Screening sensitivity for detection of malignant calcifications is low. Improving the detection of these early signs of cancer is important, because the majority of lesions with detectable calcifications that are not recalled immediately but detected as interval cancer or in the next screening round are invasive at the time of diagnosis.

6.1 Introduction

The purpose of breast cancer screening is to detect cancer as early as possible[34,35]. The earliest signs of non-palpable breast cancer are calcifications, which are usually associated with ductal carcinoma in situ (DCIS) but can also be present in invasive cancers[154]. In screening programs, between 12.7% and 41.2% of women are recalled with calcifications as the only sign of cancer[73,155–157].

The Breast Imaging Reporting and Data System (BI-RADS)[31] was designed by the American College of Radiology to standardize breast imaging reporting and to provide clarity on the interpretation of breast imaging studies. A set of guidelines is supplied in the BI-RADS atlas for the interpretation of calcifications, aiding the radiologist in distinguishing suspicious calcifications from typically benign changes, such as vascular and skin calcifications. It is recommended to recall patients with suspicious calcifications for further clinical assessment, such as a biopsy[31,158]. This can inadvertently lead to false positive outcomes, since calcifications associated with benign disease often look suspicious.

The vast majority of cancers detected by calcifications are DCIS, of which <20% are low grade[159,160]. In the discussion about the pros and cons of breast cancer screening, detection of low grade cancers is generally regarded as overdiagnosis[160,161], since the detection of these cancers does not impact mortality reduction[162]. However, it is not possible to radiologically distinguish calcifications associated with low-grade DCIS from more aggressive forms (grade II and III) in mammography, while these forms should be detected as early as possible[162–164]. Therefore, radiologists in breast cancer screening are instructed to recall all suspicious calcifications. However, in practice, especially in countries where screening programs pursue very low recall rate (i.e. the percentage of screening exams that are recalled in screening)[125], radiologists do not recall patients with calcifications without the reasonable likelihood that they represent DCIS. In such scenario, interpretation of calcifications depends more on the training, experience, and skill of the screening radiologists in a dual reading setting.

There are many studies in which screening mammograms have been retrospectively evaluated to determine the sensitivity of the screening in detecting breast cancer[67,165–175]. For instance, Vitak[167] re-examined screening exams performed prior to the diagnosis of 544 interval cancers, i.e. cancers diagnosed between screening exams usually due to symptoms, reporting that 25% of these patients could have been recalled based on the screening mammogram. Destounis et al.[173] have found that cancer was visible in 31% of 318 exams prior to a later screen detection, while Burhenne et al.[67] have shown that cancer was visible in 67% of 427 such cases. Broed-

ers et al.[172] have shown that half of 234 screen-detected and interval cancers were already visible on a prior exam. Other studies have reported that around 40% of screen-detected cancers could be detected on a previous exam [166,171,176]; however, note that all the aforementioned studies have been performed on screen-film mammography results and cannot be directly compared to digital mammography, the current acquisition standard in breast cancer screening. Several studies have shown that recall rates and cancer detection rates resulting from suspicious calcifications differ significantly between screen-film or digital mammography[73,177]. Studies by Knox et al.[174] and Weber et al.[175], in which digital mammography screening performance was assessed, have reported that between 10.5% and 31% of interval cancers were missed in screening.

In most of these studies, no distinction was made between soft tissue lesions and calcifications, and generally only interval cancers were evaluated to determine false negatives. However, cancers that were detectable but missed in a prior screening can also be considered as false negatives. In this study, we include false negatives on prior mammograms of both screen-detected and interval cancers. We focus on earlier detection of calcifications, which can prevent the development of invasive disease. A better understanding of this phenomenon is not only relevant in relation to interval cancers, but also to screen-detected cancers, independent of the threshold used for recall.

The purpose of this study is to estimate how often malignant calcifications are not detected in a population-based screening program with double reading, and to determine the proportion of invasive cancers detected by the presence of calcifications that were not recalled in the previous screening round. For this purpose, an accurate assessment of the presence of calcifications in mammograms was performed in a large screening cohort using a computer aided detection (CAD) system, in combination with visual inspection by an experienced radiologist. This provided a solid ground truth for the analysis. This also allowed us to accurately assess the sensitivity of screening for calcifications associated with breast cancer, in programs equipped with modern digital mammography systems.

6.2 Materials

All data used in this study were collected from a single region of the Dutch Breast Cancer Screening Program (Bevolkings Onderzoek Midden-West, The Netherlands). In the Dutch Breast Cancer Screening Program, women between the age of 50 and 74 are biennially invited for a screening exam. This database contained all of the available screening exams, consisting of medio-lateral oblique and cranial-caudal views

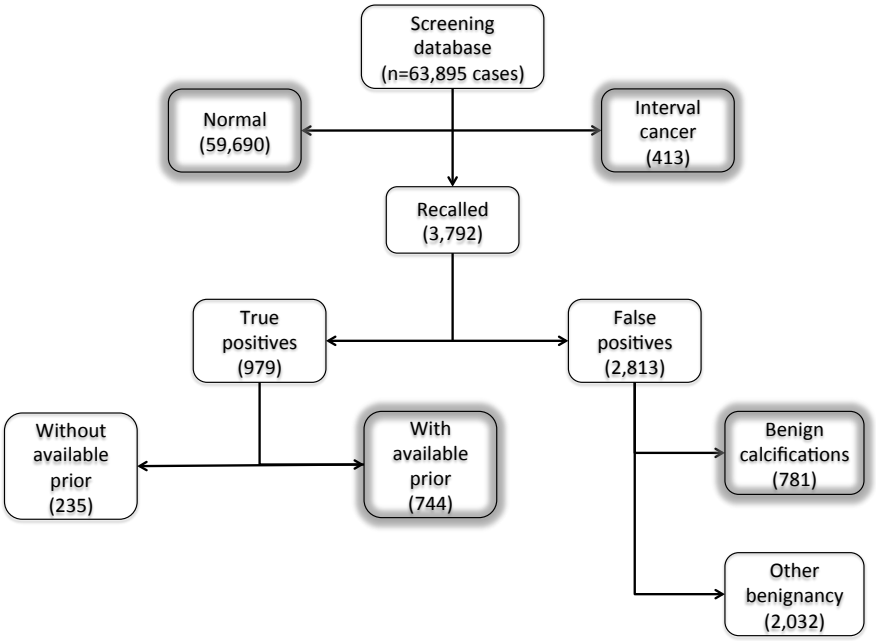


Figure 6.1: Overview of the breast cancer screening database used in this study, with data from 2003 to 2014. In this study, we include 170,878 screening exams from 63,895 women. Boxes with a grey glow were included in the ground truth for the evaluation.

of the left and right breasts, from all screened women between 2003 and 2014. All images were acquired with full-field digital mammography systems (Hologic, Bedford, Massachusetts, United States). After acquisition, two radiologists independently assessed the mammogram and scored both breasts according to the BI-RADS. When there is a discrepancy between scores, a consensus meeting is held and when no consensus is reached a third radiologist breaks the tie[40]. Women with BI-RADS 4 or 5 are recalled for further investigation.

An overview of the screening database is shown in Figure 6.1. During the study period, 63,895 women (age: 59 ± 7) participated in the screening program (with a total of 170,878 screening exams). In 59,690 women, no abnormalities were found in any of their screening exams. A total of 3,792 women were recalled for diagnostic follow up, of whom 979 had breast cancer. The remaining 2,813 recalled women were false positives, of whom 781 were recalled based on calcifications only. In 413 women, an interval cancer was found between screening exams.

6.3 Methods

To construct the ground truth, we identified all women in our database with a pathologically proven invasive and non-invasive breast cancer for whom a negative prior screening exam was available in addition to the screening mammogram that led to

the detection of the cancer. Furthermore, we identified all negative screening exams prior to interval cancers, and all exams that were recalled solely based on calcifications but were pathologically proven to be benign. Exams of women without any abnormalities were included as well. We excluded women who were recalled multiple times (n=49) or those who were recalled and later diagnosed with an interval cancer (n=11), to avoid complications in the analysis. The data included in this study are highlighted in Figure 6.1.

To determine the ground truth regarding the presence of mammographically visible calcifications associated with cancer, we used radiology and pathology reports and a retrospective review of negative prior exams by a radiologist with more than 25 years of experience in reading mammograms and more than 15 years certified as a screening radiologist. To reduce the subjectivity and workload of the radiologist, a state-of-the-art CAD system[74,178], operating at its highest sensitivity, was first applied to all mammograms of women with a screen-detected or an interval cancer. The CAD system was developed in house, but we took care that mammograms used to train the system were not included in the study dataset to avoid bias. This training set comprised less than 1% of the total number of normal exams in the screening database and less than 3% of the screen-detected cancers with a prior exam. Before the radiologist inspected the cases, an initial visual inspection by a researcher with experience in reading mammograms was carried out to exclude false positive CAD findings that were obviously not related to the recalled malignancies such as detected noise or vascular calcifications. The radiologist visually inspected the remaining exams to determine whether they were related to the later diagnosis of screen-detected or interval cancer. Prior exams of screen-detected cancers were visually inspected together with the subsequent screening mammogram and radiology reports in which the cancer was detected. For the interval cancers, diagnostic mammograms and radiology reports were not available because the anonymized data in the database could not be linked to the hospitals where the assessment took place. Only the laterality of the interval cancer was known. The visual assessment was performed on a 12MP Coronis Uniti mammography monitor (Barco N.V., Kortrijk, Belgium).

In the constructed database, the number of exams with detectable calcifications was determined for the false negative exams, which contained visible calcifications related to the cancer prior to the diagnosis of a screen-detected cancer (n_{prior}) or an interval cancer ($n_{interval}$). The number of exams with detectable calcifications was also determined for the true positive screening exams, which did not have visible calcifications in the prior exam (n_{SD}). In this way, each woman with malignant calcifications was represented only once in the series. The screening sensitivity for de-

tecting calcifications associated with cancer was calculated as follows:

$$Sensitivity = \frac{n_{SD}}{n_{SD} + n_{prior} + n_{interval}} \tag{6.1}$$

The proportion of invasive and non-invasive cancers at the time of detection was calculated to assess how often women with calcifications at the site of the detected cancer were diagnosed with invasive cancers. The invasive status of each cancer was obtained from the pathology reports. Finally, to analyze tumor size at the time of detection, the tumor stage (as T1, T2, or T3[179]) was also collected for all invasive cancers.

6.4 Results

Exams of all 744 screen-detected cancers and 1,157 exams obtained prior to a screen-detected or interval cancer were processed with the CAD system. In 536 of the 1,157 prior exams, the CAD system detected at least one instance of calcifications. Of these, the researcher classified 112 as obvious false positives that were not related to the cancer. CAD findings in the remaining 434 exams were visually inspected by the radiologist, who determined that 177 exams contained calcifications related to cancer. Figure 6.2 shows three examples of non-recalled screening exams with calcifications: 1) prior to a screen-detected cancer with calcifications, 2) prior to a soft tissue lesion, and 3) prior to an interval cancer.

By including the calcifications detectable in prior exams, we identified 325 exams with calcifications associated with malignancy in our dataset. Of these exams, 45.5% (n_{SD} =148) had calcifications that were only detectable at the time of the recall. The remaining 54.5% ($n_{prior} + n_{interval}$ =177) were detectable in the previous negative mammograms: 36.3% (n_{prior} =118) on exams prior to a positive screening exam and 18.2% ($n_{interval}$ =59) on exams prior to an interval cancer. An overview of the distribution of these prior exams is shown in Table 6.1. These numbers were used to compute the sensitivity for the detection of calcifications associated with breast cancer in digital mammogram screening; the screening sensitivity for malignant calcifications was calculated to be 45.5%. The specificity of malignant calcification detection, which was calculated from 166,673 exams without abnormalities and 781 false positive exams with calcifications, was 99.5%. This means that only 0.5% of the exams were falsely recalled in screening based on calcifications alone. Table 6.2 summarizes the invasive status of the cancers. Of the 148 screen-detected cancers detected with calcifications, but without visible calcifications on the prior exam, 77 (52.4%) were invasive. Of the screen-detected cancers with calcifications visible in the prior exams, 71 (60.2%) were invasive when they were detected in the following screening round

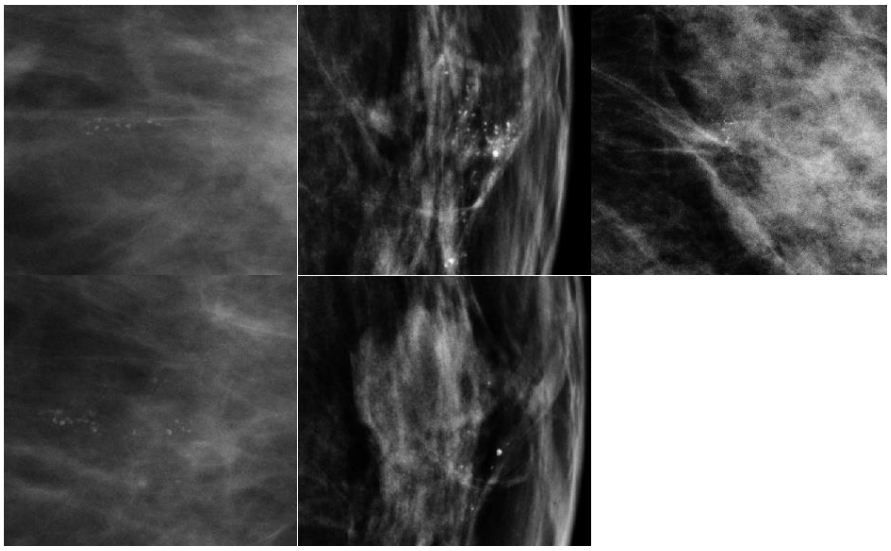


Figure 6.2: Examples of calcifications detectable on prior mammograms. The top row contains examples of exams prior to screen-detected cancer with calcifications (first column), prior to a soft tissue lesion (second column), and prior to an interval cancer (third column). In each exam, a radiologist identified detectable calcifications related to the cancer. In the bottom row, the same locations are shown as above, focusing on where the cancer and soft tissue lesions were detected (first and second column, respectively).

Assessment of calcifications related to cancer in prior exams			
	Prior exams available	**Prior exams assessed by radiologist**	**Prior exams with calcifications related to cancer**
Prior to screen-detected cancer	744	222	118
Prior to interval cancer	413	212	59
Total number of prior exams	*1,157*	*434*	*177*

Table 6.1: Overview of the prior exams included in the ground truth. A CAD system was applied to each prior exam. Only the exams with calcifications found in the same region as the cancer were visually assessed by a radiologist, who identified calcifications related to cancer.

Distribution of invasive and non-invasive cancers			
	Total number of detectable cancers	**Invasive**	**Non-invasive**
Screen-detected calcifications	148	77 (52.4%)	71 (47.6%)
Earlier-detectable calcifications associated with cancer, prior to:			
Screen-detected malignancies	118	71 (60.2%)	47 (39.8%)
Interval cancers	59	50 (84.7%)	9 (15.3%)
Total earlier-detectable cancers	*177*	*121 (68.4%)*	*58 (31.6%)*

Table 6.2: Distribution of invasive and non-invasive cancers for all screen-detected calcifications and for all cancers with calcifications detectable in exams prior to screen-detected or interval cancer diagnosis.

(i.e. two years later). Of the interval cancers with visible calcifications in the negative prior mammogram, 50 (84.7%) were invasive once they were detected. Overall, of all the 177 detectable calcifications associated with cancer from the prior exams, 121 (68.4%) developed into an invasive disease. The tumor stage for all invasive cancers with detectable calcifications is shown in Table 6.3. Of the screen-detected cancers detected by calcifications with negative prior exams, 22.1% were stage T2 or T3. For the invasive cancers detectable by calcifications on the exam prior to a recall or prior to an interval cancer, the percentage at tumor stage T2 and T3 were 25.3% and 40.0%, respectively. For all cancers that could have been detected earlier from their associated calcifications, the percentage of invasive cancers at tumor stage T2 or T3 was 31.4%. The tumor stage was not available for 11 invasive cancers.

6.5 Discussion and Conclusion

In this study, we determined the sensitivity of a population-based screening program with double reading for calcifications associated with breast cancer using digital mammography. By considering all detectable malignant calcifications visible in exams prior to a screen-detected or an interval cancer diagnosis, we found that the screening sensitivity for malignant calcifications in the studied program was only 45.5%, while the specificity for calcifications was 99.5%. Because double reading is practiced in the screening program, we believe that it is unlikely that the generally

Distribution of tumor stages					
	Total number of invasive cancers	**Stage T1 (<2cm)**	**Stage T2 (2-5cm)**	**Stage T3 (>5cm)**	**Unknown**
Screen-detected calcifications	77	59 (76.6%)	15 (19.5%)	2 (2.6%)	1 (1.3%)
Earlier-detectable calcifications associated with cancer, prior to:					
Screen-detected malignancies	71	51 (71.8%)	17 (23.9%)	1 (1.4%))	2 (2.8%)
Interval cancers	50	22 (44.0%)	18 (36.0%)	2 (4.0%)	8 (16.0%))
Total earlier-detectable cancers	*121*	*73 (60.3%)*	*35 (28.9%)*	*3 (2.5%)*	*10 (8.3%)*

Table 6.3: Distribution of the tumor stages for invasive cancers with calcifications.

low sensitivity is caused by the oversight of the radiologists; instead, it is more likely that these results reflect a high threshold in the judgment of the radiologists when characterizing calcifications as suspicious or unimportant.

This high threshold for recalling calcifications as a strategy to minimize overdiagnosis should perhaps be revised. We found that 68.4% of the women with cancer who had calcifications in a negative prior screening mammogram had developed an invasive cancer by the time it was detected. This could suggest that lowering of the threshold for recall in the national screening program is justifiable because more invasive cancers could be detected earlier. The frequency of invasive disease in women recalled with calcifications that were not detectable in prior images was 52.4%.

This finding indicates that in screening programs with a low recall rate, earlier detection of the calcifications visible in prior exams might prevent up to 16% of cancers from becoming invasive. Earlier detection would also reduce the occurrence of more advanced cancers; 31.4% of the invasive cancers with calcifications detectable on a prior exam presented as a stage T2 or T3 disease, i.e. the cancer was larger than 20 mm, at the time of diagnosis, compared to 22% when no calcifications were present in the prior exam.

In this study, we found that 54.5% of the screen-detected and interval cancers were detectable by calcifications in exams prior to diagnosis. In previous studies

it was found that 3167% of screen-detected cancers could have been identified in earlier mammograms and 1031% of the interval cancers were visible in prior screenings[67,165–175]. In our study, the percentage of cancer-related calcifications detectable on an exam prior to the later diagnosis of screen-detected cancers was 44%, and 14% of the exams prior to interval cancer diagnoses contained detectable calcifications related to the cancer. These percentages are within the ranges reported in literature; however, since previous studies did not make a distinction between calcifications and soft tissue lesions when calculating the number of cancers that could have been detected earlier, the results cannot be compared directly. Moreover, most previous studies were performed with screen-film mammography rather than digital mammography, which can have a different effect on the recall and cancer detection rates, especially for recalls based on calcifications[73,177].

We constructed the ground truth by applying a CAD system for the detection of calcifications in the prior mammograms, followed by visual inspection by an experienced radiologist to determine presence of calcifications related to the cancer detected later. The main purpose of using CAD was to reduce the workload of the radiologist and to make his judgment more objective. Because the CAD system was very sensitive, one could argue that use of a CAD system in screening could improve detection; however, current commercial CAD systems only provide mark regions for further attention to avoid calcifications being overlooked. For the detection of all calcifications in the ground truth, the specificity of the CAD system was only 51% when applied to the whole screening database and for the setting at which it was used. It should be noted that this specificity was achieved by considering all CAD marks irrespective of their scores; therefore, increasing the threshold on these scores could increase the specificity of the CAD system but reduce its maximum sensitivity. While this setting may be appropriate for use of CAD as a perception aid, it leaves the difficult problem of deciding which women with calcifications the radiologists should recall. To increase the role of CAD in calcification characterization algorithms, these systems should be developed to find an acceptable balance between sensitivity and specificity that would best help radiologists to stratify calcifications by risk. Previous studies have already demonstrated that CAD algorithms outperform radiologists in this task and there is potential to improve them considerably using new machine learning techniques[74,132,151,178,180–184].

A limitation of this study is that we do not know how many negative exams did contain calcifications. Negative exams will contain many benign calcifications and most likely also some malignant calcifications that did not yet result in a diagnosis of cancer within the two-year follow-up period we used for verification. Sometimes, benign calcifications are categorized as BI-RADS 2, but they are not always reported.

It would be interesting to study how often benign calcifications occur that look suspicious but were not recalled, and to compare them to the malignant calcifications that were missed in screening. However, in this study visual assessment of the large number of negative exams was not performed. Therefore, we cannot assess to what extent a higher recall of suspicious calcifications would lead to a strong increase of false positives.

Another limitation of our study is that we did not have access to information of all interval cancers in the period between 2013 and 2014. The absence of these cases and the exclusion of 60 cases with multiple recalls may have had a small effect on the results we present. The missing interval cancer information, as well as the absence of the radiology reports for interval cancers, can only lead to an underestimation of the number of detectable malignant calcifications and, due to this, the reported sensitivity may be slightly overestimated. Another limitation of our study is that it is based on data from one Dutch screening center, which may not be representative of other breast cancer screening programs. In particular, the radiologists in the center operated at a low recall rate, following the Dutch national breast cancer screening policy. Within Europe, the recall rate varies from 2% to 6%[125], with the screening program in the Netherlands operating at a recall rate of around 2.5%. In the United States, recall rates are substantially higher[126]. It is noted, however, that the interval cancer rate in the Dutch program and the percentage of cancers visible on prior mammograms are similar to those reported in the literatures[67,165–175,185]. This shows that our study data is representative of other screening practices.

To conclude, 54.5% of calcifications associated with cancer could potentially be detected earlier and this may substantially reduce the occurrence of invasive cancers in the screened population. It is therefore important to develop techniques that allow the earlier recall of patients with calcifications without increasing false positives and invasive diagnostic procedures to unacceptable levels.

Performance of a standalone CAD system and 109 radiologists

7

Jan-Jurre Mordang, M. Kallenberg, Albert Gubern-Mérida, Thijs Kooi, Mireille Broeders, Gerard den Heeten, and Nico Karssemeijer

Original title: Comparing the performance of a computer system for automated reading of mammograms with 109 screening radiologists

Published in: To be submitted

Abstract

Purpose: To compare the diagnostic performance of an automated Deep Learning based (DL) system for detection of breast cancer in mammograms to that of 109 certified screening radiologists using data from a national self-assessment test.

Methods: Breast cancer detection performance of 109 Dutch screening radiologists was evaluated with a self-assessment test in 2012. This test contains 60 mammography cases including 35 cases without any abnormalities and 25 cases with a screen-detected cancer. Cases were selected from a cohort in the Dutch breast cancer screening program by an expert panel consisting of three radiologists. An in-house developed DL system was applied to this dataset, resulting in a suspiciousness score for each case. The performance of the DL system was evaluated with Receiver-Operating Characteristics (ROC) analysis and was compared to the performance of the individual radiologists.

Results: The average sensitivity of the radiologists was 82.1% ($\pm$13.6%) at a specificity of 92.2% ($\pm$7.3%). The area under the ROC curve for the average radiologist was not significantly different than the automated detection system (respectively, 0.91 versus 0.89, p=0.35). However, the majority of the radiologists (95 out of 109, 87.2%) performed better than the computer system.

Conclusions: An automated system for detecting breast cancer in mammograms performs not significantly different the average of 109 certified screening radiologists on the study dataset. However, this data set only contains 60 cases and validation on a larger study is necessary.

7.1 Introduction

Early detection of breast cancer leads to a reduction in breast cancer mortality[34,35]. Therefore, breast cancer screening programs have been implemented in many developed countries. In these programs, women between the age of 50 and 70 are periodically invited for a screening mammogram, where age range and screening interval are depending on the screening policy which differs per country[186]. In most European screening programs, reading of these mammograms is performed with double reading where two radiologists read each mammogram and score each breast side. In the United States, double reading is not commonly performed. However, Computer-aided Detection (CAD) systems are widely used in screening practice in the United States[67,101]. These systems consist of computerized algorithms that can fully automatically analyze each mammogram. After analysis, suspicious locations in the mammograms are marked by the system. These marks can be shown on request by the radiologists during reading sessions.

The benefit of including CAD systems into breast cancer screening is that they can prevent oversight errors and can help the radiologist to differentiate between malignant lesions and benign lesions that looks suspicious[61,62,187,188]. Therefore, in the past decades, these systems have been under development and are still a prominent research topic. Several reader studies have shown that using a CAD system to support the reading of mammograms can increase the performance of individual radiologists[73,94,95]. However, there is little evidence that current routine use of CAD in breast cancer screening results in a significant improvement of the overall breast cancer screening performance[96,98,101].

Existing CAD systems have been developed as an aid for radiologists to avoid overlooking suspicious abnormalities. Due to limitations of algorithm performance these systems have many false positives, which make them less suitable for use as second reader. However, with the recent progress in machine learning better systems can be developed, which operate at the level of an experienced radiologist. Such systems have potential to serve as an independent reader. The incorporation of CAD as an independent observer can be done in several ways. For instance, CAD can be used as a pre-selection tool to only select women where an abnormality is found for double reading while cases without any found abnormalities are only shown to a single radiologist. Another implementation of a standalone CAD system is to use it as a replacement of one of the two radiologists in double reading or by using the CAD system as an additional independent reader in screening programs with single reading.

Several studies have been performed on the evaluation of a standalone CAD sys-

tem in screening mammography[65,67,88,94,189–194]. In three of these studies, performed by Hupse et al (2013)[190], by Kooi et al (2016)[88,195], and Becker et al[194], results of CAD were directly compared to the performance of individual radiologists. In the study of Hupse et al[190], the performance of a CAD system, aimed at detecting soft-tissue lesions, was compared to the performance of 9 radiologists and 3 residents. Their results showed that the CAD performance was similar to the performance of certified radiologists at a specificity of 0.95. The studies performed by Kooi et al[88,192] were also only performed with a soft-tissue lesion detection system. Their results showed that when only regions of interest are considered a deep learning system outperformed the average of 4 radiologists in classifying malignant and non-malignant regions. A limitation of both studies is that calcifications were not included while these are findings in screening as well. Suspicious calcifications are present in around one third of all screen-detected breast cancers in digital mammography[73,155–157]. Hence, the inclusion of calcifications is needed for an accurate evaluation of a CAD system as an independent observer in breast cancer screening. In a study by Becker et al[194], a deep learning approach has been evaluated and compared to three radiologists[194]. This study showed that their system was comparable to the radiologists. However, the evaluation of this study was limited to few radiologist and not done on data obtained from breast cancer screening. Furthermore, radiological characteristics of the lesions included in their datasets were not published.

The purpose of this study is to evaluate a CAD system based on deep learning, which detects both soft-tissue lesions and calcifications, as an independent observer on population based screening data by comparing it to the performance of screening radiologists. Evaluation is performed with data from a self-assessment test for Dutch breast screening radiologists, in which 109 certified screening radiologists participated.

7.2 Materials

The performance of individual radiologists can be evaluated with a self-test (also known as a proficiency test or self-assessment test) where radiologists read a set of cases, often enriched with breast cancers[196–201]. In 2012, the Dutch Expert Centre for Screening unrolled a self-test for all registered Dutch screening radiologists[201].

The mammograms included in this self-test, as well as the performance measures for each radiologist, were used in this study. The self-test dataset was composed by an expert panel of three radiologists who were engaged in educational activities and audits at the NETCB. Each of the radiologists within the expert panel had more than 10 years of experience in reading screening mammograms. The mammograms in

Distribution of cases in the self test	
No Abnormalities	35
Soft tissue lesion	17
Calcifications	5
Soft tissue lesion + calcifications	3
Total number of cases	60

Table 7.1: Overview of the self-test data set

these sets were obtained from the Dutch Breast Cancer Screening Program Mid-West (The Netherlands) and contained the mammographic images of the medio-lateral oblique and cranial-caudal views of each breast. Additionally, the mammographic images of the prior screening round (two years before) were included when available, nine cases were initial screening examinations and therefore did not have a prior screening round. All mammographic images were acquired with full-field digital mammography systems (Hologic, Bedford, Massachusetts, United States).

The breast cancer categorization in the self-test is shown in Table 7.1. The self-test contained the exams of 60 women: 25 with breast cancer and 35 without abnormalities (i.e. normal cases). All breast cancers in the dataset were histopathologically proven. Furthermore, the radiological characteristics, i.e. soft-tissue lesion or suspicious calcifications, of the breast cancers were also available. The radiologists who completed the self-test were registered in the NETCB quality registry that requires the radiologist to read a minimum of 3,000 mammography screens per year. Reading of the self-test was performed on the same diagnostic workstations that are daily used for breast cancer screening (Hologic SecurView DX; Hologic, Bedford, Massachusetts, United States). During reading, radiologists gave a BI-RADS score of 0, 1, 2, 4, or 5 in an online reporting system (Ziltron software; Ziltron, Dublin, Ireland). The sensitivity and specificity were calculated for each radiologist based on their BI-RADS scores. True positives were defined when a case with cancer received a BI-RADS score of 0, 4, or 5 (i.e. would have been recalled in a screening program). When a case with cancer received a score of 1 or 2, it was considered as a false negative. False positive cases were defined as normal cases that would have been recalled.

7.3 Methods

Similar to most breast cancer detection systems, the CAD system in this study consists of two subsystems: a soft-tissue lesion detection system and a calcification detection system, both based on deep learning[88,202]. The output of these two systems is combined to generate an overall output score per exam. The CAD system was trained with a very large database containing many exams with cancer. The exams in the self-test were not used for training and thus never seen by the system.

The soft-tissue lesion detection system consists of two steps[88]. In the first step, an image processing system is applied to find a limited set of potential soft-tissue lesion locations in every image[83]. At these locations, square regions of interest (patches) are selected and potential lesions are segmented[203]. These patches together with features describing the context, location, geometry, texture and contrast of the segmented lesions are classified into normal or malignant by a deep learning system[88]. The output of the soft-tissue lesion detection system is a region of each segmented lesion together with a suspiciousness score supplied by the deep learning algorithm.

The calcification detection system consists of three main steps. Similar to the soft-tissue lesion detection system, a sensitive detector is applied to each mammogram to obtain potential calcification locations. This detector consists of a deep learning system, which determines whether a pixel belongs to a calcification or not[202]. In the second step, detected calcifications are segmented and clustered together to form groups[74]. In the third and final step, calcification groups are classified as being benign or malignant[143]. The output of the calcification detection system is a set of calcification groups, where each group contains a suspiciousness score determined by the group classifier.

The output of the two detection systems is combined in a final step, in which also a classifier is trained to predict presence of a malignant process. In this classification stage, calcification and soft-tissue lesion findings are combined when they are less than 15 mm apart from each other. Together with their location and several exam-based features (e.g. total number of findings found in an exam or breast side), the final classifier is trained to provide an output score for the whole exam on a scale from 0-100.

7.4 Evaluation

The output scores of the automated detection system were used to determine performance using Receiver Operating Characteristics (ROC) analysis. To compare the result to that of the radiologists, we first calculated the ROC curve for each indi-

Area under the ROC curve			
	Average radiologist	**Standalone CAD system**	**p-value**
Calcification detection system	0.85 ± 0.05	0.83 ± 0.09)	0.44
Soft tissue lesion detection system	0.91 ± 0.02	0.90 ± 0.05	0.40
Standalone CAD system	0.91 ± 0.02	0.89 ± 0.05	*0.35*

Table 7.2: Area under the ROC curve for each sub-detection system and the overall CAD system.

vidual radiologist based on their BI-RADS ratings. For this purpose, the BI-RADS ratings were converted to a linear scale on which BI-RADS 0 ratings were set to 3[204]. Subsequently, the average ROC curve of the radiologists was calculated and fitted with a binomial distribution with R version 3.4.3 (R Foundation for Statistical Computing, Vienna, Austria). Results were calculated for each sub-system individually (i.e. the calcification detection system and soft-tissue lesion detection system) and for the combined system. Additionally, the area under the curve (AUC) was calculated for both ROC curves calculated on the whole data set. Statistical analysis was performed with bootstrapping[90] to take variance due to the limited size of the case sample into account, with p $<$0.05 considered to show significance.

7.5 Results

All cases in the self-test were successfully processed by the CAD system. Both subsystems, i.e. for detecting calcifications and for detecting soft-tissue lesions, detected all breast cancers. The resulting AUC values with the confidence intervals are shown for each individual subsystem in Table 7.2. This Table shows that both the calcification detection system achieved a similar AUC (0.83 ± 0.09) compared to the average radiologist (0.85 ± 0.05, p=0.44) and the soft-tissue lesion detection system, 0.90 ± 0.05 versus 0.91 ± 0.02 (p = 0.40) for the standalone CAD system and radiologist, respectively. Furthermore, the overall CAD system also achieved a similar AUC for the standalone CAD system compared to the average radiologist, 0.89 ± 0.05 versus 0.91 ± 0.02 (p = 0.35), respectively. Furthermore, all comparisons between the CAD system and average radiologists were not significant.

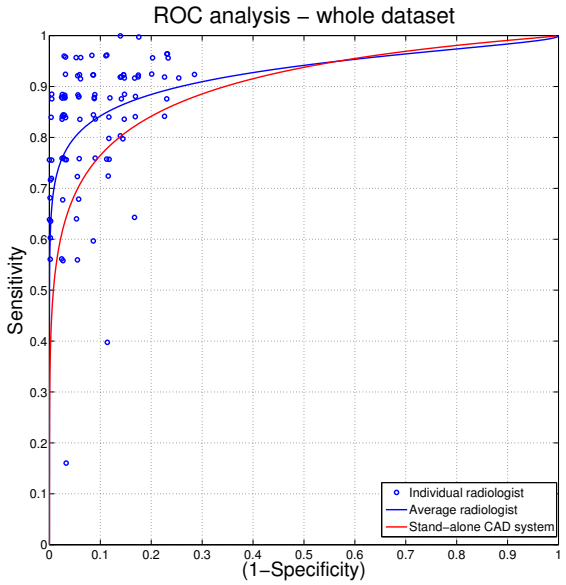


Figure 7.1: ROC plot for the automated detection system applied as a standalone CAD system. Each individual blue dot represents the performance of a radiologist. The blue line is the ROC curve of the average radiologist. The red line is the performance of the detection system on the same data set. Note that the blue dots are jittered for better visibility of overlapping points

In Figure 7.1, the ROC curve of the combined system on the self-test is shown together with the average ROC curve for screening radiologists. Additionally, the individual operating point for each individual radiologist is plotted. Note that for display some jitter is added to the individual radiologists operating points to show overlapping points. The ROC graph of the CAD system shows that the CAD system achieves an equal or better performance than 14 (12.8%) of all 109 radiologists. The average sensitivity of the radiologists was 82.1% ($\pm$ 13.6%) at an average specificity of 92.2% ($\pm$ 7.3%). The independent CAD system achieved a specificity of 79.5% at the radiologists' average sensitivity of 82.1%. The system achieved a sensitivity of 71.2% at the radiologists' average specificity of 92.2%. At 100% specificity, the sensitivity of the independent CAD system was 58.5%.

7.6 Discussion and Conclusion

In this study, we evaluated the performance of an independent automated breast cancer detection system for reading screening mammograms using data from a national self-test. This system was compared to a total of 109 radiologists and the results show that the standalone CAD system performs not significantly different than

the average radiologist. However, although there is no significant difference, the majority of the radiologists (95 radiologists, 87.2%), as well as the average radiologist, showed a better performance than the detection system.

Although the evaluated detection system in this study is getting close to the radiologists' performance, the results show that the performance of this system is still not yet optimal. Further improvements of the CAD system are necessary. An important limitation of the current CAD system is that it determines its output solely on individual mammographic views. It does not combine information from CC and MLO views nor does it make use of prior mammograms. When radiologists read mammograms, they use all information available to them such as the two different views of each breast, the differences between the left and right breast, and the difference between the current and prior mammograms. This information boosts the performance of screening radiologists[205,206]. In literature, several studies were performed to analyze the effect of including more information in the CAD system[207–217]. For example, including the information of two views for mass detection can increase the case-based sensitivity of the CAD system with 5-14% compared to only using one view for mass detection[208,209]. Furthermore, other studies have shown that including the temporal information[210–214], i.e. using prior mammograms, or including (asymmetrical) information between the left and right breast[215–217] can improve the performance of CAD systems as well. Including this information can improve the CAD system considerably. However, these studies were all performed on soft-tissue lesion CAD systems and the effect is not yet evaluated on calcification CAD systems. Therefore, further research should be done to evaluate the incorporation of multi-view and multi-exam information into the overall detection system.

Although the use of self-tests has shown to be a useful tool to evaluate the performance of individual screening radiologists[196–201], there are still several concerns. These concerns are mainly about the datasets that are used. The biggest concern is the limited number of cases (i.e. 60) in each self-test dataset, which is quite low. In other self-test datasets the number of mammograms were 50[200] and 109[218]. This number of cases is important because too many cases will yield less radiologists to participate in the self-test as this will take too much time and effort and too few cases will limit the (statistical) power of the self-test. However, the optimal number of cases for a self-test has not yet been studied[201]. Furthermore, the self-test datasets are laboratory datasets, enriched with breast cancer cases, and not representative for screening data. For example, the prevalence of abnormal cases can have an effect on the performance outcomes[219]. Consequently, it is not known how well the radiologists and the standalone CAD system will perform in practice. The question remains if the CAD system would perform better, worse, or similar to the 109 radiologists.

Compared to other studies in which the performance of a standalone CAD system was compared to the performance of radiologists[88,190,194,195], the evaluated CAD system in this study performed similar. To compare our results to the studies performed by Hupse et al[190] and Kooi et al[88,195], only the performance of the soft tissue detection system can be compared. In these studies, also, no significant difference could be found between the radiologists performance and the standalone CAD system. In the study performed by Becker et al[194], the performance of the standalone system was compared to two datasets. In one dataset their study showed a significant lower performance of their system compared to two radiologists and one of the radiologists showed almost equal performance. In the second dataset, performance of their system was not significantly different from the performance of the three radiologists.

In this study, we have shown that the performance of an independent automated deep learning system is nearing the performance of individual breast cancer screening radiologists in a small and enriched set of mammograms. The performance of the standalone CAD system was in this set not significantly different than the average radiologist computed from 109 certified screening radiologists. Further research in more daily circumstances with a much lower prevalence of pathology and comparison to double reading by radiologists should be the next step. However, in those reading programs where a second certified radiologist is not available a standalone CAD system could be a very welcome adjunct to a screening program already.

Summary

Breast cancer is the one of the most deadly types of cancer in the female population[1,2]. In the Netherlands, 1 of every 7 women develops breast cancer during her lifetime and early detection of this type of cancer can reduce breast cancer related mortality[34,35]. Breast cancer screening programs are implemented in most developed countries, and millions of mammograms are acquired each year leading to a substantial workload for radiologists, especially in screening programs with double-reading. To reduce reading time and improve detection, computer-aided detection (CAD) systems have been developed. These systems can analyze each mammographic image and localize suspicious regions. On reading workstations, these regions can be shown together with a score indicating their suspiciousness.

It has been proven that these CAD systems can improve the detection rate of individual radiologists at the cost of a slight increase in recall rate. When implemented in breast cancer screening, however, no convincing evidence has been found for a benefit of using these systems as an aid. One of the main reasons that might explain this disspoint is that because CAD systems show a large amount of false positive marks in each mammograms. These false positives arise because the CAD systems are set to operate at a very high sensitivity such that hardly any cancers are missed, reducing its specificity. This relatively high number of false positives can have a substantial impact on the general acceptance of these systems in the screening workflow. For instance, it can lead to an increase in the number of women being unnecessarily recalled[73,96,97], an increase in interpretation time of the mammograms[102], and a loss of confidence in the CAD system[102]. In practice, it is difficult to assess if CAD marks are actually judged by the radiologist during screening or generally ignored when CAD is available to them and using CAD is reimbursed[101,103].

A better solution might be to use the CAD system as a completely independent observer instead of using it as an aid for the radiologists during mammography screening. In this setup, the CAD system can serve as a stand-alone system and its output can be, for example, used as a pre-selection tool or as an additional reader in single reader or a replacement of one reader in double reading. When used as a pre-selection tool, the CAD system analyzes each mammogram and selects the cases with a detected abnormality for human reading. In this setup, the radiologists do not have to focus on filtering out the large amount of normal exams but focus on the more difficult tasks such as the differentiation of benign and malignant abnormalities. Another setup could be to use CAD as an additional reader, either as a first reader or second. Possibly even as a third reader, in the case of a disagreement.

In this thesis, we have focused on the development of a stand-alone CAD system where our general goal was to optimize this system for the detection of calcifications, the earliest signs of breast cancer in mammography.

In the first part of this thesis we mainly focus on the improvement of a calcification CAD system by reducing the number of false positives in mammograms. In **Chapter 2**, we have improved the initial detection of calcification candidates with a deep learning approach involving convolutional neural networks (CNNs). These networks are highly suitable for this task because they can overcome the large class-imbalance between the relatively few pixels belonging to calcifications and the vast amount of pixels belonging to other breast structures. Additionally, CNNs learn the most descriptive features from the data itself without the need of pre-configured ("hand-made") features and CNNs can be applied very fast to new unseen data which is a prerequisite as millions of images have to processed each year. The newly trained CNN was compared to the cascade classifier, a state-of-the-art candidate detector in mammography. This cascade classifier consists of a sequence of individual classifiers trained on Haar-like features where after each classifier samples are filtered out such that each subsequent classifier can focus on more difficult samples. On a database including individually annotated calcifications in mammograms acquired on mammography units developed by three different vendors, the results showed that the CNN significantly outperformed the cascade classifier independent of mammography unit manufacturer.

In **Chapters 3 and 4**, we have focused on the removal of one of the most frequent causes of false positives, breast arterial calcifications (BACs). Because BACs are present in a relatively small fraction of the screening population, we have proposed a framework in which we first select cases with BACs and only remove BACs in these cases as positive CAD marks. This case selection procedure is described in **Chapter 3**. In this chapter, we have trained a case-based classifier which uses features calculated on all mammographic images per case and classifies each case as either a case containing BACs or not. With this approach already 44% of the cases with BACs can be selected without selecting any case without BACs and all cases with BACs were selected a specificity of 80%. In **Chapter 4**, a method is proposed to reduce the number of false positive CAD marks due to BACs in the cases, selected by the method in the previous chapter. The BACs removal method included a novel set of features aimed at differentiating BACs from malignant calcifications. The new method was evaluated with and without the case selection and compared to a current state-of-the-art calcification CAD system. In this comparison, we found that the BACs removal increases performance of the CAD system significantly and when adding the case selection the performance of the system increased even more.

Next to BACs, other benign calcifications types, with each a different origin, can be detected by the CAD system as false positives. Several of CAD marks indicating typically benign calcifications would be easily dismissed by the radiologists as

they are obviously false positives and are only bothersome when encountered during reading. These obvious false positives (OFPs) include three different types of CAD findings: 1) BACs, 2) macrocalcifications, and 3) detection errors such as detected noise in the mammogram. Therefore, we have proposed a method to remove these obvious false positive CAD marks in mammography screening data which is described in **Chapter 5**. In this chapter, we have evaluated two approaches where in the first one the conventional classification method is used to differentiate OFPs from malignant calcifications. In the second approach, a multi-class method is proposed where individual classifiers are trained for each type of OFP together with a classifier combining their output scores. Both methods were evaluated on an independent test set and compared to the baseline CAD system. The results in this chapter show that classification of OFPs significantly improves the CAD system compared to the baseline system. Furthermore, when dedicated classifiers are trained for each of the three different OFP subtypes followed by a combinator classifier, the performance is significantly better than the conventional two-class classification strategy.

In the second part of this thesis, we have evaluated the potential of CAD for implementation in breast cancer screening. First, as described in **Chapter 6**, we have evaluated the breast cancer screening program sensitivity in detecting calcifications and the importance of detecting calcification early. In this chapter, we evaluated a large breast cancer screening cohort containing the digital mammograms of 63,895 individual women who participated in the breast cancer screening program from 2003 until 2014. In this database, we have assessed the number of calcifications that are detectable on a mammogram prior to a screen-detected or interval cancer. The ground truth was established by applying a calcification CAD system to all these prior mammograms and inspection of a radiologist of CAD findings to remove false positives of CAD and to identify calcifications retrospectively related to cancer. In this retrospective analysis, it was found that almost 55% of the detectable calcifications are visible in a prior mammogram. Furthermore, the majority (68.4%) of these detectable calcifications became an invasive cancer (rather than remaining DCIS). These results suggest that sensitivity of detection of calcifications in screening should be improved.

In **Chapter 7**, we have evaluated a complete stand-alone CAD system, including both a mass CAD and calcification CAD system developed in this thesis. In this evaluation, the performance of the CAD system was determined on a dataset, which is also used to evaluate screening radiologists of the Dutch breast cancer screening program in a self-test. This self-test was carried out in 2012 and the performance results of a total of 109 radiologists were available to us. The results in this chapter show that the CAD system performance is not significantly different than the

performance of screening radiologists although the majority of the radiologists, as well as the average radiologist, outperform the standalone CAD system. However, in breast cancer screening programs where double reading is not common or where a second certified radiologist is not available a stand-alone CAD system could be a very welcome adjunct to a screening program.

General discussion

Calcification CAD systems

In this thesis, several methods are described that significantly improved an existing calcification CAD system when compared to a current state-of-the-art system. This baseline system consists of three stages, where each of these stages can have a noticeable effect on each other and are equally important. For instance, the initial candidate detection, where each pixel in the image is classified as belonging to a calcification or not, determines the maximum sensitivity of finding calcifications. In the subsequent step, calcifications are segmented and grouped together. In the final stage, all groups are classified to remove groups of calcifications that are not malignant, i.e. contain benign types of calcifications or only noise. In this thesis we have mainly focused on improving two stages of the CAD system: the (pixel-based) candidate detection step and the false positive removal step. However, the segmentation and grouping stage can be further developed as well. The general focus for further development of this stage could be in making this stage more robust. At the moment, segmentation is sensitive to the threshold that is set for defining individual calcifications such that modifying the candidate detector stage can result in changes in the number of segmented calcifications and, consequently, clusters. The biggest concern in this stage is to find an optimal balance between detecting true calcifications and removal of detected noise. For instance, when not all calcifications are detected this can lead to a low sensitivity of the overall system. However the system should also be specific enough because more noise in the clusters makes the false positive removal more difficult. In this thesis we have focused more on the detection and characterization of groups of calcifications which has a direct influence on the performance of the CAD system. However further research in increasing the robustness of the segmentation and grouping is still recommended.

In recent years, medical image analysis has been more and more focused on using deep learning in detecting pathologies, segmenting medical images, or improving diagnosis[108–110,220]. The power of deep learning lies in its self learning abilities, which allows developers to build powerful applications without having detailed knowledge of the medical imaging problem they try to solve. However, many examples are needed for these deep learning systems to be trained successfully, e.g. thousands or millions of both positive and negative examples. In this thesis, the available amount of samples (i.e. small patches) seemed appropriate for training the calcification candidate detection system with deep learning. For this purpose, millions of negative samples were randomly extracted from hundreds of mammograms and thousands of individually annotated calcifications were taken as positive samples. More samples could further improve the detection. However, the num-

ber of mammograms containing malignant calcifications is limited and annotating individual calcifications is a time consuming and tedious task. Another approach is to perform a whole image classification, i.e. train a deep learning algorithm on the whole image instead of patches. However, more than thousands of images with malignant calcifications are needed otherwise training a deep learning system becomes difficult. For the negative class, on the other hand, still enough samples could be collected because, fortunately, there are many cases without any abnormalities in a breast cancer screening. Up until now, only few studies have managed to train a deep learning system for detection and characterizing breast cancer[221–225]. Though these studies seem promising, further research should be done to see how these systems compare to each other and to "classical" breast cancer detection methods.

When it comes to the stand-alone performance of our developed CAD system, its sensitivity would be acceptable with a sensitivity >90%. However this comes with an exam specificity of <95% (see Figure 5.8) and, consequently, a large fraction of false positives. At this operating point, approximately 5% of the screened women will be incorrectly recalled based on calcifications alone. Keep in mind that the false positives generated by the soft-tissue lesion CAD are not considered yet. Because several types of benign calcifications can be present in the breast, we have focused on the removal of (obvious) false positives detected by the calcification CAD system, first by removal of only the arterial calcifications and second by removal of other benign types of calcifications. This approach has proven to result in a significant decrease in false positives compared to the baseline system. However, there are still other types of benign calcifications that we did not yet specifically aim to remove from the output of the CAD system. These types of calcifications should still be removed when developing a stand-alone CAD system to obtain a more specific independent observer.

Other improvements of the CAD system are also necessary. An important limitation of the current CAD system is that it analyzes only one image at the time, while additional information can be derived from comparing the CC and MLO views and current and prior mammograms to boost the radiologists' performance[205,206]. Studies have already shown that when using such comparisons in CAD systems it can boost their performance as well[207–217]. However, most of these studies were done on soft-tissue lesion detection systems and its effect on calcifications should still be evaluated.

CAD systems in breast cancer screening

Detection and characterization of calcifications that develop into breast cancer are difficult tasks, for both radiologists and computers. Radiologists operate at very high specificity and it appears that very often screen-detected and interval cancers with malignant calcifications are visible on the prior screening mammogram in retrospect. The calcification CAD system developed in this thesis, on the other hand, can be set to a very high sensitivity but lacks an appropriate specificity. The relatively low specificity of the developed CAD system makes it not yet possible to be used as a fully independent reader of screening mammograms.

Nevertheless, radiologists can still benefit from the developed CAD system as a reading tool. For example during reading, CAD could be used interactively. In this reading approach, the radioligist can query CAD results for specific locations that are of interest[65]. By clicking on the region of interest, the system will show a marker on the screen together with the CAD score. This interactive approach has shown to be more effective compared to the "tradditional" method where CAD marks are prompted by the radiologist[65]. However, the "traditional" method can still be helpful to radiologists because calcifications are relatively small and easy to miss. Therefore, it could be useful to prompt calcification CAD marks to check for oversight errors. For example, calcifications groups that contain only four calcifications or less could still be missed by the radiologists. By prompting the calfications marks, the group of calcifications can be brought to light. The radiologist can then examine the detected region and, for instance, compare it to the prior exam to see if it is a new group of calcifications. For this purpose, a suitable balance should be found between what the CAD system shows and what the radiologist expect to see, i.e. a more sensitive CAD system will show more false positives.

Although in this work, a large part of the false positives have been removed as resulting findings of the developed calcification CAD system, there is still work to do to further improve the quality of the system to a level where it can be accepted as an independent reader of mammograms. The most obvious addition to the system is that it should be combined with a detection system of soft-tissue lesions such that it can find all types of breast cancer. In Chapter 7, we have investigated an "overall" CAD system, but it did not yet perform as well as the average radiologist. Considering the 2-6% recall rate in European screening programs, the developed CAD system does not yet achieve a fitting sensitivity. However, the CAD system could still be used standalone. For instance, when set to its highest sensitivity (i.e. close to 100% detection of breast cancer), the specificity is around 50% (for calcifications, see Chapter 6). When we assume similar specificity for the mass CAD system, ex-

ams with a relatively high chance of containing an abnormality could be selected for double reading and double reading might not be needed for the other half of the screening exams because there is a very low probability that there will be any abnormalities in these. This approach could decrease the workload of the radiologists theoretically by 25% or increase the time they have to read the suspicious cases.

Future prospects

Step-by-step breast cancer detection systems are becoming as good as screening radiologists. With the current research trend in artificial intelligence and deep learning, the performance of these cancer detection systems will increase even more and more and if this trend is continuous, eventually, pass the performance of most experience radiologists. This, however, does not mean that radiologists will be replaced by these systems. In the end, whatever work flow changes there will be and how much perception tasks will be shifted to a CAD system, a radiologist has to decide if recall is justified. At the moment, radiologists have the ultimate medical responsibility as this is required by law. At this stage, the CAD systems could already be used as one of the two readers in double reading or as a selection tool for double or single reading. Moreover, when the system has been validated on large screening datasets, it might even be contemplated to only read the screening exams selected by the CAD system. It is the right time for large studies in high volume circumstances like the screening as it is organized in a number of European countries.

Samenvatting

Borstkanker is een van de meest dodelijke kankers in de vrouwelijke populatie[1,2]. In Nederland heeft een op de zeven vrouwen kans om borstkanker te ontwikkelen tijdens haar leven en vroege detectie van deze kanker reduceert de mortaliteit[34,35]. Daarom zijn er borstkankerscreeningprogramma's geïmplementeerd in de meeste ontwikkelde landen. Omdat een groot deel van de vrouwelijke populatie is uitgenodigd voor dit bevolkingsonderzoek worden er elk jaar miljoenen mammogrammen gemaakt wat leidt tot een substantiële werkdruk voor radiologen, voornamelijk wanneer alle beelden door twee radiologen bekeken worden. Om de leestijd te verminderen en de detectie van borstkanker te vergroten zijn er computergestuurde detectiesysteem ontwikkeld. Een dergelijk systeem analyseert elk mammogram en lokaliseert verdachte gebieden. Ieder gebied krijgt uiteindelijk van dit systeem een verdachtheidsscore.

Het is bewezen dat deze computergestuurde detectiesystemen de radiologen helpen meer kankers te vinden ten koste van een kleine toename in het aantal doorverwijzingen. Echter, wanneer het systeem geïntegreerd is in de borstkankerscreening is er geen overtuigend bewijs gevonden dat deze systemen een nuttige toevoeging zijn in de algehele detectie van borstkanker. Een van de grootste nadelen van de huidige (commerciële) systemen ontstaat doordat deze systemen een hoge sensitiviteit moeten waarborgen en dat gaat ten koste van de specificiteit, wat leidt tot een grote hoeveelheid foutpositieve markeringen in elk mammogram. Deze grote hoeveelheid foutpositieven kan een sterk effect hebben op de algemene acceptatie van dit soort systemen in het bevolkingsonderzoek. Dit kan bijvoorbeeld leiden tot een toename van vrouwen die onnodig doorverwezen worden[73,96,97], een toename van leestijd van mammogrammen[102] en een verlies van vertrouwen in het systeem.[102]. Het is daardoor moeilijk om vast te stellen of de markeringen van het systeem ook echt bekeken worden door de radioloog tijdens de screening of dat ze over het algemeen genegeerd worden ook al is een detectiesysteem beschikbaar.[101,103].

Het systeem inzetten als een onafhankelijke beoordelaar van mammogrammen zou daarom een betere toepassing van het systeem kunnen zijn in plaats van als hulp voor de radioloog bij het lezen van mammogrammen. In deze configuratie wordt het systeem ingezet als een onafhankelijke waarnemer en de analyse kan bijvoorbeeld gebruikt worden als een selectiemiddel of als een extra beoordelaar in screening. Als een selectiemiddel kan het systeem na de analyse automatisch bepalen of een mammogram een abnormaliteit bevat en alleen deze worden dan doorgestuurd naar de screeningsradiologen. In deze setup hoeven radiologen niet te focussen op het wegfilteren van de "normale" mammogrammen, maar kunnen in plaats daarvan meer focussen op de moeilijkere taken zoals het differentiëren van benigne en maligne abnormaliteiten in de mammogrammen. Het systeem kan ook als een extra beoorde-

laar worden ingezet waarbij er één van de twee radiologen vervangen kan worden door het systeem of bij een meningsverschil tussen de twee radiologen kan het systeem optreden als een derde lezer.

Het huidige systeem moet echter nog een stuk verbeterd worden om dit doel te halen, met name het verminderen van het aantal foutpositieven. Daarom hebben is er in deze thesis gefocust op het ontwikkelen en verbeteren van een computergestuurd detectiesysteem in mammogrammen. Ons hoofddoel is het optimaliseren van de detectie van calcificaties, het vroegste stadium van borstkanker.

In het eerste deel van deze thesis hebben we voornamelijk gefocust op het verbeteren van het calcificatie detectiesysteem door het verminderen van het aantal foutpositieve markeringen in de mammogrammen. In **Hoofdstuk 2** is de detectie van individuele calcificaties verbeterd met een zogeheten Deep-Learning methode waarbij convolutionele neurale netwerken gebruikt worden. Deze netwerken zijn uiterst geschikt voor deze taak omdat ze ongevoelig zijn voor het grote klassegrootte verschil tussen het relatief kleine aantal pixels behorende bij een calcificatie in vergelijking met het grote aantal pixels van andere borststructuren. Een ander voordeel is dat de convolutionele neurale netwerken zelf de meeste descriptieve kenmerken uit de data kunnen halen zonder dat er extra informatie nodig is die de data beschrijven. Deze netwerken kunnen tevens snel toepast worden op nieuwe ongeclassificeerde data wat een vereiste is voor een dergelijk systeem, omdat het miljoenen mammmogrammen moet kunnen verwerken per jaar. Het nieuw getrainde systeem was vergeleken met een cascade classificatie methode. Deze methode is een state-of-the-art methode en bestaat uit een sequentie van individuele classifiers die getraind zijn met zogeheten Haar-like features. In de cascade wordt na iedere individuele classifier pixels weggefilterd zodat de latere classifiers meer kunnen focussen op moeilijker classificeerbare pixels. De vergelijking was gedaan op een database wat individuele geannoteerde calcificaties bevat. Deze data is verkregen met mammografen van drie verschillende bedrijven. De resultaten van deze studie lieten zien dat het gebruik van convolutionele neurale netwerken de calcificatiedetectie significant verbeterde in vergelijking met de cascade classificatie, onafhankelijk van de mammograaffabrikant.

In **Hoofdstuk 3 en 4** is er gefocust op het verwijderen van een van de meest voorkomende oorzaken van foutpositieven: arteriële calcificaties in de borst. Omdat dit type calcificaties maar een een relatief klein deel van de screeningspopulatie voorkomt, stellen we een methode voor waarbij we eerst de vrouwen selecteren met arteriële calcificaties. En alleen in die geselecteerde vrouwen de foutpositieven te verwijderen die worden veroorzaakt door arteriële calcificatie. De selectie van vrouwen met arteriële calcificaties is beschreven in **Hoofdstuk 3**. In dit hoofdstuk is

een classifier getraind dat van iedere vrouw alle mammogrammen van een screeningsronde analyseert en gebaseerd op alle informatie in de mammogrammen bepaalt of dat er arteriële calcificaties gevonden zijn. Met deze methode konden we al 44% van de vrouwen met arteriële calcificaties selecteren zonder per ongeluk een vrouw te selecteren zonder dit type calcificaties. Alle vrouwen met arteriële calcificaties konden met dit systeem gevonden worden, waarbij 20% van de vrouwen werd geselecteerd zonder deze calcificaties. In **Hoofdstuk 4** is een methode beschreven dat het aantal arteriële calcificaties verwijderde als positieve markeringen in de mammogrammen van vrouwen die werden geselecteerd door de methode beschreven in het voorgaande hoofdstuk. Voor de methode om arteriële calcificaties te verwijderen hebben we een nieuwe set van kenmerken ontworpen die de arteriële calcificaties zoveel mogelijk differentieert van maligne calcificaties. De nieuwe methode was geëvalueerd met en zonder de selectie van vrouwen met arteriële calcificaties en vergeleken met de huidige state-of-the-art calcificatiedetectiemethode. In de evaluatie hebben we aangetoond dat het verwijderen van arteriële calcificaties leidt tot een significante verbetering van het systeem. Het voorselecteren van vrouwen met arteriële calcificaties en door alleen de verwijderingsstap te doen in de mammogrammen van deze vrouwen verbeterde het systeem nog meer.

Naast arteriële calcificaties bevinden zich ook andere benigne soorten calcificaties in de borst met elk een andere oorsprong. Het detecteren van deze calcificaties wordt ook gezien als foutpositieven van het detectiesysteem. Een aantal van deze foutpositieven zijn typisch benigne en worden eenvoudig herkend door de radioloog en genegeerd. Het markeren van deze duidelijke foutpositieven (DFP) zullen alleen maar als vervelend worden beschouwd door radiologen. De groep DFPs bevat drie verschillende typen foutpositieven: 1) arteriële calcificaties, 2) macro-calcificaties (calcificaties groter dan 10mm) en 3) detectiefouten zoals ruis in het mammogram. In **Hoofdstuk 5** is een methode beschreven dat deze DFPs verwijdert als detectiemarkeringen van het detectiesysteem. In dit hoofdstuk hebben we twee methoden beschreven en geëvalueerd. Hierbij was de eerste methode de conventionele manier van classificatie is toegepast voor het differentiëren van DFPs en maligne calcificaties. De tweede methode bestaat uit een multi-classificatie waarbij individuele classifiers zijn getraind op ieder DFP type en gecombineerd zijn om de DFP's te verwijderen als detectiemarkeringen. Beide methodes zijn geëvalueerd op een onafhankelijke dataset en vergeleken met de huidige state-of-the-art calcificatiedetectiemethode. De resultaten in deze studie toonden aan dat het verwijderen van DFP's het detectiesysteem significant verbeterde. De tweede methode, de multi-classificatie methode, presteerde zelfs significant beter dan de conventionele classificatiemethode.

In het tweede deel van deze thesis hebben we de potentie van CAD in borstkankerscreening. Allereerst hebben we in **Hoofdstuk 6** het belang van het vroegtijdig detecteren van maligne calcificaties in de borstkankerscreening en hebben we de sensitiviteit in het detecteren van maligne calcificaties van het borstkankerscreeningprogramma bepaald. In dit hoofdstuk hebben we groot borstkankerscreeningcohort geëvalueerd. Dit cohort bevatte digitale mammogrammen van 63,895 individuele vrouwen die hebben deelgenomen aan de borstkankerscreening tussen 2003 en 2014. In deze databases hebben we gekeken naar het aantal calcificaties die detecteerbaar zijn op het mammogram voorafgaande aan de screeningsronde waarbij een borstkanker was gevonden of een intervalkanker (een kanker die ontwikkeld tussen twee screeningsronden). Om dit te doen hebben we met een calcificatiedetectiesysteem alle voorgaande mammogrammen geanalyseerd en de mammogrammen waar het systeem iets vond laten inspecteren door een radioloog. De radioloog selecteerde hierbij alleen de mammogrammen waarbij het detectiesysteem de later gevonden kanker heeft gemarkeerd. Hierbij vonden we dat bijna 55% van detecteerbare calcificaties zichtbaar zijn in het voorgaande mammogram (twee jaar eerder). Bovendien ontwikkelde het merendeel van deze eerder detecteerbare calcificaties tot een invasieve kanker wat het belang van het vroeg detecteren in borst kankerscreening versterkt.

In **Hoofdstuk 7** hebben we een compleet zelfstandig detectiesysteem geëvalueerd. Dit systeem detecteerde tumor schaduwen en calcificaties. Met dit detectiesysteem hebben we een zelftestdataset geanalyseerd en de prestatie van het zelfstandige detectiesysteem was vergeleken met screenings radiologen. Deze dataset is gebruikt voor het evalueren van screenings radiologen en de Nederlandse borstkankerscreening. Deze zelftest was uitgevoerd in 2012 en de resultaten van 109 radiologen waren beschikbaar voor ons. In deze studie hebben we laten zien dat het detectiesysteem niet significant verschillend presteert dan de screeningsradiologen.

Publications

Papers in international journals

T. Tan, **J.J. Mordang**, J. van Zelst, A. Grivegnée, A Gubern-Mérida, J. Melendez, R.M. Mann, W. Zhang, B. Platel, and N. Karssemeijer. "Computer-aided detection of breast cancers using Haar-like features in automated 3D breast ultrasound". *Medical Physics*, 42(4) 1498-1504, 2015.

J.J. Mordang, A. Gubern-Mérida, G.J. Den Heeten, and N. Karssemeijer. "Reducing false positives of microcalcification detection systems by removal of breast arterial calcifications". *Medical Physics*, 43(4) 1676-1687, 2016.

J.J. Mordang, A. Gubern-Mérida, A. Bria, F. Tortorella, G.J. Den Heeten, and N. Karssemeijer. "Improving computer-aided detection assistance in breast cancer screening by removal of obviously false-positive findings". *Medical Physics*, 44(4) 1390-1401, 2017.

J.J. Mordang, A. Gubern-Mérida, A. Bria, F. Tortorella, R.M. Mann, M. Broeders, G.J. Den Heeten, and N. Karssemeijer. "The importance of early detection of calcifications associated with breast cancer in screening". *Breast Cancer Research and Treatment*, 1-8, 2017

M.T. Oei, F.J. Meijer, **J.J. Mordang**, E.J. Smit, A.J. Idema, B.M. Goraj, H.O. Laue, M. Prokop and R. Manniesing. "Observer Variability of Reference Tissue Selection for Relative Cerebral Blood Volume Measurements in Glioma Patients". *European Radiology*, 28(9) 39023911, 2018.

A. Bria, C. Marrocco, L. Borges, M. Molinara, A. Marchesi, **J.J. Mordang**, N. Karssemeijer, F. Torterella. "Improving the Automated Detection of Calcifications using Adaptive Variance Stabilization". *IEEE Transactions on Medical Imaging*, 37(8) 1857 - 1864, 2018.

A. Rodriguez-Ruiz, E. Krupinski, **J.J. Mordang**, K. Schilling, S.H. Heywang-Kobrunner, I. Sechopoulos, R.M. Mann. "Detection of breast cancer using mammography: Impact of an Artificial Intelligence support system". *Radiology*, in press, 2018.

Papers in conference proceedings

J.J. Mordang, M.T.H. Oei, R. van den Boom, E.J. Smit, M. Prokop, B. van Ginneken, and R. Manniesing. "A pattern recognition framework for vessel segmentation in 4D CT of the brain". In: *Medical Imaging*, volume 8669 of Proceedings of the SPIE, 2013.

J.J. Mordang, J. Hauth, G.J. den Heeten, and N. Karssemeijer. "Automated Labeling of Screening Mammograms with Arterial Calcifications". In: *Breast Imaging*, volume 8539 of the 13th workshop on Digital Mammography, 2014.

J.J. Mordang and N. Karssemeijer. "Vessel segmentation in screening mammograms". In: *Medical Imaging*, volume 9414 of Proceedings of the SPIE, 2015.

J.J. Mordang, T. Janssen, A. Bria, T. Kooi, A. Gubern-Mérida, and N. Karssemeijer. "Automatic microcalcification detection in multi-vendor mammography using convolutional neural networks". In: *Breast Imaging*, volume 9699 of the 13th workshop on Digital Mammography, 2016.

T. Kooi, A. Gubern-Mérida, **J.J. Mordang**, R.M. Mann, R. Pijnappel, K. Schuur, G.J. den Heeten, and N. Karssemeijer. "A comparison between a deep convolutional neural network and radiologists for classifying regions of interest in mammography". In: *Breast Imaging*, volume 9699 of the 13th workshop on Digital Mammography, 2016.

A. Bria, C. Marrocco, **J.J. Mordang**, N. Karssemeijer, M. Molinara, and F. Tortorella. "LUT-QNE: Look-up-table quantum noise equalization in digital mammograms". In: *Breast Imaging*, volume 9699 of the 13th workshop on Digital Mammography, 2016.

T. Kooi, **J.J. Mordang**, and N. Karssemeijer. "Conditional random field modelling of interactions between findings in mammography". In: *Medical Imaging*, volume 10134 of Proceedings of the SPIE, 2017.

A Bria, C. Marrocco, A. Galdran, A. Campilho, A. Marchesi, **J.J. Mordang**, N. Karssemeijer, M. Molinara, and F. Tortorella. "Spatial Enhancement by Dehazing for Detection of Microcalcifications with Convolutional Nets". In: *Image Analysis and Processing*, volume 10485 of the 19th International Conference on Image Analysis and Processing, 2017.

A. Marchesi, A. Bria, C. Marrocco, M. Molinara, **J.J. Mordang**, F. Tortorella, N. Karssemeijer. "The effect of mammogram preprocessing on microcalcification detection with convolutional neural networks". In: **CBMS**, IEEE 30th International Symposium on Computer-Based Medical Systems (CBMS), 2017

A. Rodriguez-Ruiz, **J.J. Mordang**, N. Karssemeijer, I. Sechopoulos, R. Mann. "Can radiologists improve their breast cancer detection in mammography when using a deep learning-based computer system as decision support?". In *Breast Imaging*, 14th

International Workshop on Breast Imaging, volume 10718 of Proceedings of the SPIE, 2018

B. Savelli, A. Bria, C. Marrocco, M. Molinara, **J.J. Mordang**, N. Karssemeijer, F. Tortorella, "Improving the automated detection of calcifications by combining deep cascades and deep convolutional nets". In *Breast Imaging*, 14th International Workshop on Breast Imaging, volume 10718 of Proceedings of the SPIE, 2018

C. Marrocco, A. Bria, V. Di Sano, L.R. Borges, M. Molinara, **J.J. Mordang**, N. Karssemeijer, F. Tortorella, "Mammogram denoising to improve the calcification detection performance of convolutional nets". In *Breast Imaging*, 14th International Workshop on Breast Imaging, volume 10718 of Proceedings of the SPIE, 2018

Abstracts in conference proceedings

M.T.H. Oei, B.M. Goraj,F.J.A. Meijer, **J.J. Mordang**, A.J. Idema, S.H.E. Boots-Sprenger, H.O.A. Laue, and M. Prokop. "Variability of relative cerebral blood volume normalization in patients with gliomas: Interobserver and intraobserver reproducibility study", *Annual meeting of the International Society for Magnetic Resonance in Medicine*, 2011.

J.J. Mordang, M.T.H. Oei, R. van den Boom, H.O.A. Laue, L.J. Oostveen, M. Prokop, B. van Ginneken, and R. Manniesing. "Effect of Dose Reduction in CT Perfusion Scans on Cerebral Blood Flow and Volume Computed with Three Perfusion Software Packages: Analysis with a Digital Phantom". In: *Annual Meeting of the Radiological Society of North America"*, 2012.

Bibliography

[1] Jemal A., Bray F., Center M. M., Ferlay J., Ward E., and Forman D. Global cancer statistics. *CA: a cancer journal for clinicians*, 61(1542-4863):69–90, 2011.

[2] Torre L. A., Bray F., Siegel R. L., Ferlay J., Lortet-Tieulent J., and Jemal A. Global cancer statistics, 2012. *CA: a cancer journal for clinicians*, 65(2):87–108, mar 2015.

[3] IKC. Integraal kankercentrum Nederland. URL `https://www.iknl.nl/home`.

[4] KWF Kankerbestrijding. Kanker in Nederland tot 2020, 2011. URL `https://www.kwf.nl/SiteCollectionDocuments/rapport-Kanker-in-Nederland-tot-2020.pdf`.

[5] Vervoort M. M., Draisma G., Fracheboud J., van de Poll-Franse L. V., and de Koning H. J. Trends in the usage of adjuvant systemic therapy for breast cancer in the Netherlands and its effect on mortality. *British journal of cancer*, 91(2):242–7, jun 2004.

[6] Berry D. A., Cronin K. A., Plevritis S. K., Fryback D. G., Clarke L., Zelen M., Mandelblatt J. S., Yakovlev A. Y., Habbema J. D. F., and Feuer E. J. Effect of Screening and Adjuvant Therapy on Mortality from Breast Cancer. *New England Journal of Medicine*, 353(17):1784–1792, 2005.

[7] Otten J. D. M., Broeders M. J. M., Fracheboud J., Otto S. J., de Koning H. J., and Verbeek A. L. M. Impressive time-related influence of the Dutch screening programme on breast cancer incidence and mortality, 1975-2006. *International journal of cancer*, 123(8):1929–34, oct 2008.

[8] Louwman W. J., Voogd A. C., van Dijck J. A. A. M., Nieuwenhuijzen G. A. P., Ribot J., Pruijt J. F. M., and Coebergh J. W. W. On the rising trends of incidence and prognosis for breast cancer patients diagnosed 1975–2004: a long-term population-based study in southeastern Netherlands. *Cancer Causes & Control*, 19(1):97–106, 2008.

[9] Lopez-Garcia M. A., Geyer F. C., Lacroix-Triki M., Marchió C., and Reis-Filho J. S. Breast cancer precursors revisited: molecular features and progression pathways. *Histopathology*, 57(2):171–92, aug 2010.

[10] Cowell C. F., Weigelt B., Sakr R. A., Ng C. K. Y., Hicks J., King T. A., and Reis-Filho J. S. Progression from ductal carcinoma in situ to invasive breast cancer: Revisited. *Molecular Oncology*, 7(5):859–869, 2013.

[11] Bloom H. J. G. and Richardson W. W. Histological Grading and Prognosis in Breast Cancer: A Study of 1409 Cases of which 359 have been Followed for 15 Years. *British Journal of Cancer*, 11 (3):359–377, sep 1957.

[12] ELSTON C. W. and ELLIS I. O. pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up. *Histopathology*, 19(5):403–410, 1991.

[13] Gradishar W. J., Anderson B. O., Balassanian R., Blair S. L., Burstein H. J., Cyr A., Elias A. D., Farrar W. B., Forero A., Giordano S. H., Goetz M., Goldstein L. J., Hudis C. A., Isakoff S. J., Marcom P. K., Mayer I. A., McCormick B., Moran M., Patel S. A., Pierce L. J., Reed E. C., Salerno K. E., Schwartzberg L. S., Smith K. L., Smith M. L., Soliman H., Somlo G., Telli M., Ward J. H., Shead D. A., and Kumar R. Invasive Breast Cancer Version 1.2016, NCCN Clinical Practice Guidelines in Oncology. *Journal of the National Comprehensive Cancer Network*, 14(3):324–354, mar 2016.

[14] Maas A. H. E. M., van der Schouw Y. T., Beijerinck D., Deurenberg J. J. M., Mali W. P. T. M., and van der Graaf Y. Arterial calcifications seen on mammograms: cardiovascular risk factors,

pregnancy, and lactation. *Radiology*, 240(1):33–38, 2006.

[15] Maas A. H. E. M., van der Schouw Y. T., Atsma F., Beijerinck D., Deurenberg J. J. M., Mali W. P. T. M., and van der Graaf Y. Breast arterial calcifications are correlated with subsequent development of coronary artery calcifications, but their aetiology is predominantly different. *European Journal of Radiology*, 63(3):396–400, 2007.

[16] Maas A. H. E. M., van der Schouw Y. T., Mali W. P. T. M., and van der Graaf Y. Prevalence and determinants of breast arterial calcium in women at high risk of cardiovascular disease. *The American journal of cardiology*, 94(5):655–9, 2004.

[17] Hendriks E. J. E., De Jong P. A., van der Graaf Y., Mali W. P. T. M., van der Schouw Y. T., and Beulens J. W. J. Breast arterial calcifications: A systematic review and meta-analysis of their determinants and their association with cardiovascular events. *Atherosclerosis*, 239(1):11–20, 2015.

[18] Schnatz P. F., Marakovits K. A., and O'Sullivan D. M. The association of breast arterial calcification and coronary heart disease. *Obstet Gynecol*, 117:233–241, 2011.

[19] Kataoka M., Warren R., Luben R., Camus J., Denton E., Sala E., Day N., and Khaw K.-T. How predictive is breast arterial calcification of cardiovascular disease and risk factors when found at screening mammography? *AJR. American journal of roentgenology*, 187(1):73–80, 2006.

[20] Topal U., Kaderli A., Topal N. B., Özdemir B., Yeilbursa D., Cordan J., Ediz B., and Aydinlar A. Relationship between the arterial calcification detected in mammography and coronary artery disease. *European Journal of Radiology*, 63(3):391–395, 2007.

[21] Dale P. S., Richards M., and Mackie G. C. Vascular calcifications on screening mammography identify women with increased risk of coronary artery disease and diabetes. *American Journal of Surgery*, 196(4):537–540, 2008.

[22] Rotter M. a., Schnatz P. F., Currier A. a., and O'Sullivan D. M. Breast arterial calcifications (BACs) found on screening mammography and their association with cardiovascular disease. *Menopause (New York, N.Y.)*, 15(2):276–81, 2008.

[23] Ahn K. J., Kim Y. J., Cho H. J., Yim H. W., Kang B. J., Kim S. H., Kim H. S., Kim K. T., Lee J. H., and Whang I. Y. Correlation between breast arterial calcification detected on mammography and cerebral artery disease. *Archives of Gynecology and Obstetrics*, 284(4):957–964, 2011.

[24] Bae M. J., Lee S. Y., Kim Y. J., Lee J. G., Jeong D. W., Yi Y. H., Cho Y. H., Choi E. J., and Choo K. S. Association of breast arterial calcifications, metabolic syndrome, and the 10-year coronary heart disease risk: a cross-sectional case-control study. *Journal of women's health (2002)*, 22(7): 625–630, 2013.

[25] Shah N., Chainani V., Delafontaine P., Abdo A., Lafferty J., and Abi Rafeh N. Mammographically detectable breast arterial calcification and atherosclerosis. *Cardiology in review*, 22(2):69–78, 2014.

[26] Çetin M., Çetin R., Tamer N., and Kelekçi S. Breast arterial calcifications associated with diabetes and hypertension. *Journal of Diabetes and its Complications*, 18(6):363–366, 2004.

[27] Çetin M., Çetin R., and Tamer N. Prevalence of breast arterial calcification in hypertensive patients. *Clinical Radiology*, 59(1):92–95, 2004.

[28] Reddy J., Bilezikian J. P., Smith S. J., and Mosca L. Reduced bone mineral density is associated with breast arterial calcification. *The Journal of clinical endocrinology and metabolism*, 93(1):208–211, 2008.

[29] Duhn V., D'Orsi E. T., Johnson S., D'Orsi C. J., Adams A. L., and O'Neill W. C. Breast arterial calcification: a marker of medial vascular calcification in chronic kidney disease. *Clin J Am Soc Nephrol*, 6:377–382, 2011.

[30] Zafar A. N., Khan S., and Zafar S. N. Factors associated with breast arterial calcification on mammography. *Journal of the College of Physicians and Surgeons–Pakistan : JCPSP*, 23(3):178–81, 2013.

[31] D'Orsi C. J., E.A.Sickles, Mendelson E. B., and Et al. E. A. M. *ACR BI-RADS Atlas, Breast Imaging Reporting and Data System*. 2013.

[32] Hofvind S., Iversen B. F., Eriksen L., Styr B. M., Kjellevold K., and Kurz K. D. Mammographic morphology and distribution of calcifications in ductal carcinoma in situ diagnosed in organized screening. *Acta radiologica (Stockholm, Sweden : 1987)*, 52(5):481–7, 2011.

[33] Cox R. F., Hernandez-Santana a., Ramdass S., McMahon G., Harmey J. H., and Morgan M. P. Microcalcifications in breast cancer: novel insights into the molecular mechanism and functional consequence of mammary mineralisation. *British journal of cancer*, 106(3):525–37, 2012.

[34] Tabár L., Gad A., Holmberg L., Ljungquist U., Fagerberg C., Baldetorp L., Gröntoft O., Lundström B., Månson J., Eklund G., Day N., and Pettersson F. Reduction in Mortality From Breast Cancer After Mass Screening With Mammography. *The Lancet*, 325(8433):829–832, 1985.

[35] Tabár L., Vitak B., Chen T. H.-H., Yen A. M.-F., Cohen A., Tot T., Chiu S. Y.-H., Chen S. L.-S., Fann J. C.-Y., Rosell J., Fohlin H., Smith R. A., and Duffy S. W. Swedish Two-County Trial: Impact of Mammographic Screening on Breast Cancer Mortality during 3 Decades. *Radiology*, 260(3):658–663, 2011.

[36] Sickles E. A. Mammographic features of 300 consecutive nonpalpable breast cancers. *American Journal of Roentgenology*, 146(4):661–663, 1986.

[37] Page D. L., Dupont W. D., Rogers L. W., Jensen R. A., and Schuyler P. A. Continued local recurrence of carcinoma 1525 years after a diagnosis of low grade ductal carcinoma in situ of the breast treated only by biopsy. *Cancer*, 76(7):1197–1200, 1995.

[38] Sanders M. E., Schuyler P. A., Dupont W. D., and Page D. L. The natural history of low-grade ductal carcinoma in situ of the breast in women treated by biopsy only revealed over 30 years of long-term follow-up. *Cancer*, 103(12):2481–4, jun 2005.

[39] Ferlay J., Shin H. R., Bray F., Forman D., Mathers C., and Parkin D. M. Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. *International Journal of Cancer*, 127(12):2893–2917, 2010.

[40] Holland R., Rijken H., and Hendriks J. The Dutch Population-Based Mammography Screening: 30-Year Experience. *Breast Care*, 2:12–18, 2007.

[41] RIVM. Uitvoeringskader Bevolkingsonderzoek Borstkanker. Technical report, 2015. URL `http://rivm.nl/Documenten{_}en{_}publicaties/Professioneel{_}Praktisch/Richtlijnen/Preventie{_}Ziekte{_}Zorg/Borstkankerscreening/Uitvoeringskader{_}Bevolkingsonderzoek{_}Borstkanker`.

[42] Timmers J. M. H., van Doorne-Nagtegaal H. J., Zonderland H. M., van Tinteren H., Visser O., Verbeek A. L. M., den Heeten G. J., and Broeders M. J. M. The Breast Imaging Reporting and Data System (BI-RADS) in the Dutch breast cancer screening programme: its role as an assessment and stratification tool. *European Radiology*, 22(8):1717–1723, 2012.

[43] Rubin G. D. Data explosion: the challenge of multidetector-row {CT}. 36:74–80, 2000.

[44] Herron J. and Reynolds J. H. Trends in the on-call workload of radiologists. *Clinical Radiology*, 61(1):91–96, 2006.

[45] Smith-Bindman R., Miglioretti D. L., and Larson E. B. Rising use of diagnostic medical imaging in a large integrated health system. *Health Affairs*, 27(6):1491–1502, 2008.

[46] Iglesias J. E. and Karssemeijer N. Robust initial detection of landmarks in film-screen mammograms using multiple ffdm atlases. *IEEE Transactions on Medical Imaging*, 28(11):1815–1824, 2009.

[47] Levin D. C., Rao V. M., Parker L., and Frangos A. J. Analysis of radiologists' imaging workload trends by place of service. *Journal of the American College of Radiology*, 10(10):760–763, 2013.

[48] of Radiology A. C. Lung CT Screening Reporting and Data System (Lung-RADS), 2014. URL `http://www.acr.org/Quality-Safety/Resources/LungRADS`.

[49] Weinreb J. C., Barentsz J. O., Choyke P. L., Cornud F., Haider M. A., Macura K. J., Margolis D., Schnall M. D., Shtern F., Tempany C. M., Thoeny H. C., and Verma S. PI-RADS Prostate Imaging - Reporting and Data System: 2015, Version 2. *European urology*, 69(1):16–40, jan 2016.

[50] Elmore J. G., Wells C. K., Lee C. H., Howard D. H., and Feinstein A. R. Variability in Radiologists' Interpretations of Mammograms. *New England Journal of Medicine*, 331(22):1493–1499, 1994.

[51] Beam C. A., M L. P., and Sullivan D. C. Variability in the Interpretation of Screening Mammograms by US Radiologists: Findings From a National Sample. *Arch Intern Med*, 156(2):209–213, 1996.

[52] Elmore J. G., Jackson S. L., Abraham L., Miglioretti D. L., Carney P. a., Geller B. M., Yankaskas B. C., Kerlikowske K., Rosenberg R. D., Sickles E. a., and Buist D. S. M. Variability in Interpretive Performance at Screening Mammography and Associated with Accuracy 1 Purpose : Methods : Results : Conclusion :. 253(3):641–651, 2009.

[53] Duijm L. E. M., Louwman M. W. J., Groenewoud J. H., van de Poll-Franse L. V., Fracheboud J., and Coebergh J. W. Inter-observer variability in mammography screening and effect of type and number of readers on screening outcome. *British journal of cancer*, 100(6):901–907, mar 2009.

[54] Thurfjell E. L., Lernevall K. A., and Taube A. A. Benefit of independent double reading in a population-based mammography screening program. *Radiology*, 191(1):241–4, 1994.

[55] Brown J., Bryan S., and Warren R. Mammography screening: an incremental cost effectiveness analysis of double versus single reading of mammograms. *Bmj*, 312(7034):809–812, 1996.

[56] Hofvind S., Geller B. M., Rosenberg R. D., and Skaane P. Screening-detected Breast Cancers: Discordant Independent Double Reading in a Population-based Screening Program. *Radiology*, 253(3):652–660, 2009.

[57] Doi K. Current status and future potential of computer-aided diagnosis in medical imaging.

British Journal of Radiology, 78(SPEC. ISS.):S3—-S19, 2005.

[58] Doi K. Computer-Aided Diagnosis in Medical Imaging: Historical Review, Current Status and Future Potential. *Comput Med Imaging Graph*, 31(4-5):198–211, 2007.

[59] van Ginneken B., Schaefer-Prokop C. M., and Prokop M. Computer-aided diagnosis: how to move from the laboratory to the clinic. *Radiology*, 261(3):719–32, 2011.

[60] Dodd L. E., Wagner R. F., Armato S. G., McNitt-Gray M. F., Beiden S., Chan H.-P., Gur D., McLennan G., Metz C. E., Petrick N., Sahiner B., and Sayre J. Assessment methodologies and statistical issues for computer-aided diagnosis of lung nodules in computed tomography. *Academic Radiology*, 11(4):462–475, 2004.

[61] Tang J., Rangayyan R., Xu J., El Naqa I., and Yang Y. Computer-aided detection and diagnosis of breast cancer with mammography: Recent advances. *IEEE transactions on information technology in biomedicine*, 13(2):236–251, 2009.

[62] Elter M. and Horsch A. CADx of mammographic masses and clustered microcalcifications: a review. *Medical physics*, 36(6):2052–2068, 2009.

[63] Zhu Y., Williams S., and Zwiggelaar R. Computer technology in detection and staging of prostate carcinoma: A review. *Medical Image Analysis*, 10(2):178–199, 2006.

[64] Hupse R. and Karssemeijer N. The effect of feature selection methods on computer-aided detection of masses in mammograms. *Physics in medicine and biology*, 55(10):2893–2904, 2010.

[65] Hupse R., Samulski M., Lobbes M. B., Mann R. M., Mus R., den Heeten G. J., Beijerinck D., Pijnappel R. M., Boetes C., and Karssemeijer N. Computer-aided Detection of Masses at Mammography: Interactive Decision Support versus Prompts. *Radiology*, 266(1):123–129, 2013.

[66] Birdwell R. L. The preponderance of evidence supports computer-aided detection for screening mammography. *Radiology*, 253(1):9–16, 2009.

[67] Warren Burhenne L. J., Wood S. a., D'Orsi C. J., Feig S. a., Kopans D. B., O'Shaughnessy K. F., Sickles E. a., Tabar L., Vyborny C. J., and Castellino R. a. Potential contribution of computer-aided detection to the sensitivity of screening mammography. *Radiology*, 215(2):554–562, 2000.

[68] Domingues I. and Cardoso J. S. Using Bayesian surprise to detect calcifications in mammogram images. In *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC 2014*, pages 1091–1094, 2014. ISBN 9781424479290.

[69] Helena G., Miranda B., and Cezar J. Author ' s Accepted Manuscript Computer-Aided Diagnosis System Based on Fuzzy Logic for Breast Cancer Categorization. *Computers in Biology and Medicine*, 2014.

[70] Mohamed H., Mabrouk M. S., and Sharawy A. Computer aided detection system for micro calcifications in digital mammograms. *Computer Methods and Programs in Biomedicine*, 116(3): 226–235, 2014.

[71] Apostolopoulos G., Koutras A., Christoyianni I., and Dermatas E. Computer aided classification of mammographic tissue using shapelets and support vector machines. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 8445 LNCS, pages 510–520. Springer, 2014. ISBN 9783319070636.

[72] Pereira D. C., Ramos R. P., and do Nascimento M. Z. Segmentation and detection of breast

cancer in mammograms combining wavelet analysis and genetic algorithm. *Computer Methods and Programs in Biomedicine*, 114(1):88–101, 2014.

[73] Karssemeijer N., Bluekens A. M., Beijerinck D., Deurenberg J. J., Beekman M., Visser R., van Engen R., Bartels-Kortland A., and Broeders M. J. Breast cancer screening results 5 years after introduction of digital mammography in a population-based screening program. *Radiology*, 253 (2):353–358, 2009.

[74] Bria A., Karssemeijer N., and Tortorella F. Learning from unbalanced data: A cascade-based approach for detecting clustered microcalcifications. *Medical Image Analysis*, 18(2):241–252, 2014.

[75] Karssemeijer N. Automated classification of parenchymal patterns in mammograms. *Physics in medicine and biology*, 43(2):365–378, 1998.

[76] Bick U. and Diekmann F. Digital Mammography. 234:236, 2010.

[77] Viola P. and Jones M. Rapid object detection using a boosted cascade of simple features. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition*, volume 1, pages I–511 – I–518, 2001.

[78] Torralba a., Murphy K., and Freeman W. Sharing Visual Features for Multiclass and Multiview Object Detection Detecting multiple classes / views of objects in clutter. *Pattern Analysis and Machine Intelligence*, 29(5):854–869, 2007.

[79] Papageorgiou C., Oren M., and Poggio T. A general framework for object detection. In *Sixth International Conference on Computer Vision*, pages 555–562, 1998. ISBN VO -.

[80] Lienhart R. and Maydt J. An extended set of Haar-like features for rapid object detection. In *Proceedings. International Conference on Image Processing*, volume 1, pages 0–3, 2002. ISBN 0-7803-7622-6.

[81] Veldkamp W. J. H. and Karssemeijer N. Improved method for detection of microcalcification clusters in digital mammograms. volume 3661, pages 512–522, 1999.

[82] Brzakovic D., Luo X. M., and Brzakovic U. An approach to automated detection of tumors in mammograms. *IEEE Transactions on Medical Imaging*, 9(3):233–241, 1990.

[83] Karssemeijer N. and Te Brake G. M. Detection of stellate distortions in mammograms. *IEEE Transactions on Medical Imaging*, 15(5):611–619, 1996.

[84] Sahiner B., Chan H. P., Petrick N., Helvie M. a., and Hadjiiski L. M. Improvement of mammographic mass characterization using spiculation meausures and morphological features. *Medical physics*, 28(7):1455–1465, 2001.

[85] Eltonsy N. H., Tourassi G. D., and Elmaghraby A. S. A concentric morphology model for the detection of masses in mammography. *IEEE Transactions on Medical Imaging*, 26(6):880–889, 2007.

[86] Velikova M., Samulski M., Lucas P. J. F., and Karssemeijer N. Improved mammographic CAD performance using multi-view information: A bayesian network framework. *Belgian/Netherlands Artificial Intelligence Conference*, 54:379–380, 2009.

[87] Choi J. Y., Kim D. H., Plataniotis K. N., and Ro Y. M. Computer-aided detection (CAD) of breast masses in mammography: combined detection and ensemble classification. *Physics in medicine and biology*, 59(14):3697–719, 2014.

[88] Kooi T., Litjens G., van Ginneken B., Gubern-Mérida A., Sánchez C. I., Mann R., den Heeten A., and Karssemeijer N. Large scale deep learning for computer aided detection of mammographic lesions. *Medical Image Analysis*, 35:303–312, 2017.

[89] Timp S. and Karssemeijer N. A new 2D segmentation method based on dynamic programming applied to computer aided detection in mammography. *Medical physics*, 31(2004):958–971, 2004.

[90] Efron B. and Tibshirani R. J. *An Introduction to the Bootstrap*, volume 57. CRC press, 1993. ISBN 0412042312.

[91] Bornefalk H. Estimation and comparison of CAD system performance in clinical settings. *Academic Radiology*, 12(6):687–694, 2005.

[92] Samulski M. *Classification of {B}reast {L}esions in {D}igital {M}ammograms*. PhD thesis, Radboud University Nijmegen, 2006.

[93] Samulski M., Karssemeijer N., Lucas P., and Groot P. Classification of mammographic masses using support vector machines and Bayesian networks. In Giger M. L. and Karssemeijer N., editors, *Proceedings of SPIE*, volume 6514, pages 65141J–65141J–11, 2007.

[94] Cole E. B., Zhang Z., Marques H. S., Nishikawa R. M., Hendrick R. E., Yaffe M. J., Padungchaichote W., Kuzmiak C., Chayakulkheeree J., Conant E. F., Fajardo L. L., Baum J., Gatsonis C., and Pisano E. Assessing the stand-alone sensitivity of computer-aided detection with cancer cases from the digital mammographic imaging screening trial. *American Journal of Roentgenology*, 199 (3):W392—-W401, 2012.

[95] Nishikawa R. M. Current status and future directions of computer-aided diagnosis in mammography. *Computerized Medical Imaging and Graphics*, 31(4-5):224–235, 2007.

[96] Eadie L. H., Taylor P., and Gibson A. P. A systematic review of computer-assisted diagnosis in diagnostic cancer imaging. *European Journal of Radiology*, 81(1):e70—-e76, 2012.

[97] Cupples T. E., Cunningham J. E., and Reynolds J. C. Impact of computer-aided detection in a regional screening mammography program. *American Journal of Roentgenology*, 185(4):944–950, 2005.

[98] Fenton J. J., Taplin S. H., Carney P. A., Abraham L., Sickles E. A., D'Orsi C., Berns E. A., Cutter G., Hendrick R. E., Barlow W. E., and Elmore J. G. Influence of computer-aided detection on performance of screening mammography. *The New England Journal of Medicine*, 356(14):1399–1409, 2007.

[99] Taylor P. and Potts H. W. W. Computer aids and human second reading as interventions in screening mammography: Two systematic reviews to compare effects on cancer detection and recall rate. *European Journal of Cancer*, 44(6):798–807, 2008.

[100] Bargalló X., Santamaría G., Del Amo M., Arguis P., Ríos J., Grau J., Burrel M., Cores E., and Velasco M. Single reading with computer-aided detection performed by selected radiologists in a breast cancer screening program. *European Journal of Radiology*, 83(11):2019–2023, 2014.

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

[102] Philpotts L. E. Can computer-aided detection be detrimental to mammographic interpretation?

Radiology, 253(1):17–22, 2009.

[103] Rao V. M., Levin D. C., Parker L., Cavanaugh B., Frangos A. J., and Sunshine J. H. How widely is computer-aided detection used in screening and diagnostic mammography? *Journal of the American College of Radiology*, 7(10):802–805, 2010.

[104] LeCun Y., Bengio Y., Hinton G., Y. L., Y. B., and G. H. Deep learning. *Nature*, 521(7553):436–444, 2015.

[105] Schmidhuber J. Deep Learning in neural networks: An overview. *Neural Networks*, 61:85–117, 2015.

[106] Krizhevsky A., Sutskever I., and Hinton G. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25*, pages 1097–1105, 2012.

[107] He K., Zhang X., Ren S., and Sun J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *Proceedings of the IEEE International Conference on Computer Vision*, 11-18-Dece:1026–1034, 2016.

[108] Cirean D. C., Giusti A., Gambardella L. M., and Schmidhuber J. Mitosis detection in breast cancer histology images with deep neural networks. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 8150 LNCS, pages 411–418, 2013. ISBN 9783642407628.

[109] Cruz-Roa A. A., Arevalo Ovalle J. E., Madabhushi A., and González Osorio F. A. A deep learning architecture for image representation, visual interpretability and automated basal-cell carcinoma cancer detection. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 8150 LNCS, pages 403–410, 2013. ISBN 9783642407628.

[110] Guo Y., Wu G., Commander L. A., Szary S., Jewells V., Lin W., and Shen D. Segmenting hippocampus from infant brains by sparse patch matching with deep-learned features. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 8674 LNCS, pages 308–315, 2014. ISBN 9783319104690.

[111] Simonyan K. and Zisserman A. Very deep convolutional networks for large-scale image recognition. *arXiv:14091556*, 1409.

[112] Srivastava N., Hinton G. E., Krizhevsky A., Sutskever I., and Salakhutdinov R. Dropout : A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research (JMLR)*, 15(1):1929–1958, 2014.

[113] Bornefalk H. and Hermansson A. B. On the comparison of FROC curves in mammography CAD systems. *Med Phys*, 32(2):412–417, 2005.

[114] Samuelson F. and Petrick N. Comparing Image Detection Algorithms Using Resampling. In *3rd IEEE International Symposium on Biomedical Imaging: Macro to Nano, 2006.*, pages 1312–1315, 2006. ISBN 0-7803-9576-X.

[115] Li Q., Sone S., and Doi K. Selective enhancement filters for nodules, vessels, and airway walls in two- and three-dimensional CT scans. *Medical physics*, 30(8):2040–2051, 2003.

[116] Kemmeren J. M., van Noord P. a., Beijerinck D., Fracheboud J., Banga J. D., and van der Graaf Y.

Arterial calcification found on breast cancer screening mammograms and cardiovascular mortality in women: The DOM Project. Doorlopend Onderzoek Morbiditeit en Mortaliteit. *American journal of epidemiology*, 147(4):333–341, 1998.

[117] Reddy J., Son H., Smith S. J., Paultre F., and Mosca L. Prevalence of breast arterial calcifications in an ethnically diverse population of women. *Annals of Epidemiology*, 15(5):344–350, 2005.

[118] van Noord P. A., Beijerinck D., Kemmeren J. M., and van der Graaf Y. Mammograms may convey more than breast cancer risk: breast arterial calcification and arterio-sclerotic related diseases in women of the DOM cohort. *European journal of cancer prevention : the official journal of the European Cancer Prevention Organisation (ECP)*, 5(6):483–487, 1996.

[119] Crystal P., Crystal E., Leor J., Friger M., Katzinovitch G., and Strano S. Breast artery calcium on routine mammography as a potential marker for increased risk of cardiovascular disease. *American Journal of Cardiology*, 86(2):216–217, jan 2000.

[120] Cheng J. Z., Cole E. B., Pisano E. D., and Shen D. Detection of arterial calcification in mammograms by random walks. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 5636 LNCS:713–724, 2009.

[121] Ge J., Chan H. P., Sahiner B., Zhou C., Helvie M. A., Wei J., Hadjiiski L. M., Zhang Y., Wu Y. T., and Shi J. Automated detection of breast vascular calcification on full-field digital mammograms - art. no. 691517. In *Medical Imaging 2008: Computer-Aided Diagnosis, Pts 1 and 2*, volume 6915, page 91517. International Society for Optics and Photonics, 2008. ISBN 0277-786X.

[122] Ferlay J., Steliarova-Foucher E., Lortet-Tieulent J., Rosso S., Coebergh J. W. W., Comber H., Forman D., and Bray F. Cancer incidence and mortality patterns in Europe: Estimates for 40 countries in 2012. *European Journal of Cancer*, 49(6):1374–1403, 2013.

[123] Mendelson E. B., Baum J. K., Berg W., Merritt C., and Rubin E. *Breast Imaging Reporting and Data System*, volume 40. Reston, VA, 4 edition, 2003.

[124] Nishikawa R. M. and Gur D. CADe for early detection of breast cancer-current status and why we need to continue to explore new approaches. *Academic Radiology*, 21(10):1320–1321, 2014.

[125] Van Luijt P. A., Fracheboud J., Heijnsdijk E. A. M., Den Heeten G. J., and De Koning H. J. Nation-wide data on screening performance during the transition to digital mammography: Observations in 6 million screens. *European Journal of Cancer*, 49(16):3517–3525, 2013.

[126] Smith-Bindman R., Chu P. W., Miglioretti D. L., and coll. E. Comparison of screening mammography in the United States and the United kingdom. *Jama*, 290(16):2129–37, 2003.

[127] Mordang J.-J., Hauth J., den Heeten G., and Karssemeijer N. Automated Labeling of Screening Mammograms with Arterial Calcifications. In Fujita H., Hara T., and Muramatsu C., editors, *Breast Imaging*, volume 8539 of *Lecture Notes in Computer Science*. Springer International Publishing, 2014.

[128] Mordang J. J. and Karssemeijer N. Vessel segmentation in screening mammograms. In *SPIE Medical Imaging*, volume 9414, page 94140J, 2015. ISBN 9781628415049.

[129] Shi J., Sahiner B., Chan H.-P., Ge J., Hadjiiski L., Helvie M. A., Nees A., Wu Y.-T., Wei J., Zhou C., Zhang Y., and Cui J. Characterization of mammographic masses based on level set segmentation with new image features and patient information. *Medical physics*, 35(1):280–90, 2008.

[130] El-Naqa I., Yang Y., Wernick M. N., Galatsanos N. P., and Nishikawa R. M. A support vector machine approach for detection of microcalcifications. *IEEE Transactions on Medical Imaging*, 21 (12):1552–1563, 2002.

[131] Jing H., Yang Y., and Nishikawa R. M. Detection of clustered microcalcifications using spatial point process modeling. *Physics in medicine and biology*, 56(1):1–17, 2011.

[132] Jiang Y., Nishikawa R. M., Wolverton D. E., Metz C. E., Giger M. L., Schmidt R. A., Vyborny C. J., and Doi K. Malignant and benign clustered microcalcifications: automated feature analysis and classification. *Radiology*, 198(3):671–678, 1996.

[133] Wei J., Sahiner B., Hadjiiski L. M., Chan H.-P., Petrick N., Helvie M. A., Roubidoux M. A., Ge J., and Zhou C. Computer-aided detection of breast masses on full field digital mammograms. *Medical physics*, 32(9):2827–38, 2005.

[134] McLoughlin K. J., Bones P. J., and Karssemeijer N. Noise Equalization for Detection of Microcalcification Clusters in Direct Digital Mammogram Images. *IEEE Transactions on Medical Imaging*, 23(3):313–320, 2004.

[135] van Schie G. and Karssemeijer N. Detection of Microcalcifications Using a Nonuniform Noise Model. In *IWDM '08: Proceedings of the 9th international workshop on Digital Mammography*, pages 378–384, Berlin, Heidelberg, 2008. Springer-Verlag. ISBN 978-3-540-70537-6.

[136] Sklansky J. Finding the convex hull of a simple polygon. *Pattern Recognition Letters*, 1(2):79–83, 1982.

[137] Diestel R. *Graph Theory*, volume 85 of *Graduate Texts in Mathematics*. Springer-Verlag, Heidelberg, fourth edition, 2001. ISBN 9783540261834.

[138] Koller T. M., Gerig G., Szekely G., and Dettwiler D. Multiscale detection of curvilinear structures in 2-D and 3-D image data. In *Proceedings of the IEEE Fifth International Conference on Computer Vision - ICCV 1995*, pages 864–869, 1995.

[139] Cheng J. Z., Chen C. M., Cole E. B., Pisano E. D., and Shen D. Automated delineation of calcified vessels in mammography by tracking with uncertainty and graphical linking techniques. *IEEE Transactions on Medical Imaging*, 31(11):2143–2155, 2012.

[140] Wada H., Kitada M., Sato K., Sasajima T., Miyokawa N., and Kuroda T. Prevalence of breast arterial calcification by mammography contributes to breast cancer. *Breast Cancer*, 19(3):266–269, 2012.

[141] Daye P. H. H., Jarolimek A. M., and S. The false-negative mammogram. *Radiographics*, 18(5): 11371154, 1998.

[142] Ridgeway G. Special Invited Paper . Additive Logistic Regression : A Statistical View of Boosting. 28(2):393–400, 2013.

[143] Mordang J.-J., Gubern-Mérida A., den Heeten G., and Karssemeijer N. Reducing false positives of microcalcification detection systems by removal of breast arterial calcifications. *Medical Physics*, 43(4):1676, mar 2016.

[144] Hu M.-K. Visual pattern recognition by moment invariants, computer methods in image analysis. {*IRE*} *Transactions on Information Theory*, 8:179–187, 1962.

[145] Frangi a., Niessen W., and A. F. Frangi W. J. N. Quantitation of vessel morphology from 3D

MRA. In Taylor C. and Colchester A., editors, *Unknown*, volume 1679, pages 358–367, Berlin, 1999. Springer. ISBN 978-3-540-66503-8, 978-3-540-48232-1.

[146] Cortes C. and Vapnik V. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[147] Breiman L. Random Forests. *Machine Learning*, 45(1):5–32, 2012.

[148] Hsu C.-W., Chang C.-C., and Lin C.-J. A practical guide to support vector classification. 2003.

[149] K. H., L.L. P., M.L. G., C.E. M., and Y. J. A scaling transformation for classifier output based on likelihood ratio: Applications to a CAD workstation for diagnosis of breast cancer. *Medical Physics*, 39(5):2787–2804, may 2012.

[150] He X., Samuelson F., Gallas B. D., Sahiner B., and Myers K. The Equivalence of a Human Observer and an Ideal Observer in Binary Diagnostic Tasks. In *Spie*, volume 8673, pages 1–8. International Society for Optics and Photonics, 2013. ISBN 9780819494474.

[151] Veldkamp W. J., Karssemeijer N., Otten J. D., and Hendriks J. H. Automated classification of clustered microcalcifications into malignant and benign types. *Medical physics*, 27(11):2600–2608, 2000.

[152] Wei L., Yang Y., Nishikawa R. M., and Jiang Y. A study on several machine-learning methods for classification of malignant and benign clustered microcalcifications. *IEEE Transactions on Medical Imaging*, 24(2):371–380, mar 2005.

[153] Elangeeran M., Ramasamy S., and Arumugam K. A novel method for benign and malignant characterization of mammographic microcalcifications employing waveatom features and circular complex valued - Extreme Learning Machine. In *IEEE ISSNIP 2014 - 2014 IEEE 9th International Conference on Intelligent Sensors, Sensor Networks and Information Processing, Conference Proceedings*, pages 1–6, 2014. ISBN 9781479928439.

[154] Stomper P. C., Geradts J., Edge S. B., and Levine E. G. Mammographic Predictors of the Presence and Size of Invasive Carcinomas Associated with Malignant Microcalcification Lesions Without a Mass. *American Journal of Roentgenology*, 181(6):1679–1684, dec 2003.

[155] Del Turco M. R., Mantellini P., Ciatto S., Bonardi R., Martinelli F., Lazzari B., and Houssami N. Full-Field Digital Versus Screen-Film Mammography: Comparative Accuracy in Concurrent Screening Cohorts. *American Journal of Roentgenology*, 189(4):860–866, oct 2007.

[156] Domingo L., Romero A., Belvis F., Sánchez M., Ferrer J., Salas D., Ibáñez J., Vega A., Ferrer F., Laso M. S., Macià F., Castells X., and Sala M. Differences in radiological patterns, tumour characteristics and diagnostic precision between digital mammography and screen-film mammography in four breast cancer screening programmes in Spain. *European Radiology*, 21(9): 2020–2028, 2011.

[157] Hambly N. M., McNicholas M. M., Phelan N., Hargaden G. C., O'Doherty A., and Flanagan F. L. Comparison of digital mammography and screen-film mammography in breast cancer screening: A review in the Irish Breast Screening Program. *American Journal of Roentgenology*, 193(4):1010–1018, oct 2009.

[158] Bijker N., Donker M., Wesseling J., den Heeten G. J., and Rutgers E. J. T. Is DCIS Breast Cancer, and How Do I Treat it? *Current Treatment Options in Oncology*, 14(1):75–87, 2013.

[159] Weigel S., Hense H. W., Heidrich J., Berkemeyer S., Heindel W., and Heidinger O. Digital

Mammography Screening: Does Age Influence the Detection Rates of Low-, Intermediate-, and High-Grade Ductal Carcinoma in Situ? *Radiology*, 278(3):707–713, oct 2015.

[160] van Luijt P. A., Heijnsdijk E. A. M., Fracheboud J., Overbeek L. I. H., Broeders M. J. M., Wesseling J., den Heeten G. J., and de Koning H. J. The distribution of ductal carcinoma in situ (DCIS) grade in 4232 women and its impact on overdiagnosis in breast cancer screening, 2016. ISSN 1465-542X. URL `http://breast-cancer-research.biomedcentral.com/articles/10.1186/s13058-016-0705-5`.

[161] Groen E. J., Elshof L. E., Visser L. L., Rutgers E. J. T., Winter-Warnars H. A., Lips E. H., and Wesseling J. Finding the balance between over- and under-treatment of ductal carcinoma in situ (DCIS). *The Breast*, 2016.

[162] Tabár L., Vitak B., Chen H. H., Duffy S. W., Yen M. F., Chiang C. F., Krusemo U. B., Tot T., and Smith R. A. The Swedish Two-County Trial twenty years later. Updated mortality results and new insights from long-term follow-up. *Radiol Clin North Am*, 38(4):625–651, 2000.

[163] Bansal G. J. and Thomas K. G. Screen-detected breast cancer: does presence of minimal signs on prior mammograms predict staging or grading of cancer? *Clinical radiology*, 66(7):605–608, jul 2011.

[164] Baker R., Rogers K. D., Shepherd N., and Stone N. New relationships between breast microcalcifications and cancer. *British Journal of Cancer*, 103(7):1034–1039, sep 2010.

[165] Bird R. E., Wallace T. W., and Yankaskas B. C. Breast Imaging Missed at Screening Mammography. *Radiology*, 184(3):613–617, 1992.

[166] van Dijck J. A., Verbeek A. L., Hendriks J. H., and Holland R. The current detectability of breast cancer in a mammographic screening program. A review of the previous mammograms of interval and screen-detected cancers. *Cancer*, 72(6):1933–1938, 1993.

[167] Vitak B. Invasive interval cancers in the Ostergötland Mammographic Screening Programme: radiological analysis. *European radiology*, 8(4):639–646, 1998.

[168] Duncan J., Shi P., Constable T., and Sinusas A. Physical and geometrical modeling for image-based recovery of left ventricular deformation. *Progress in Biophysics and Molecular Biology*, 69 (2-3):333–351, 1998.

[169] Daly C. A., Apthorp L., and Field S. Second round cancers: How many were visible on the first round of the UK National Breast Screening Programme, three years earlier? *Clinical Radiology*, 53(1):25–28, 1998.

[170] Saarenmaa I., Salminen T., Geiger U., Heikkinen P., Hyvärinen S., Isola J., Kataja V., Kokko M. L., Kokko R., Kumpulainen E., Kärkkäinen A., Pakkanen J., Peltonen P., Piironen A., Salo A., Talviala M. L., and Hakama M. The visibility of cancer on previous mammograms in retrospective review. *Clinical Radiology*, 56(1):40–43, 2001.

[171] Zheng B., Shah R., Wallace L., Hakim C., Ganott M. A., and Gur D. Computer-aided detection in mammography: An assessment of performance on current and prior images. *Academic Radiology*, 9(11):1245–1250, nov 2002.

[172] Broeders M. J. M., Onland-Moret N. C., Rijken H. J. T. M., Hendriks J. H. C. L., Verbeek A. L. M., and Holland R. Use of previous screening mammograms to identify features indicating cases that would have a possible gain in prognosis following earlier detection. *European Journal of*

Cancer, 39(12):1770–1775, 2003.

[173] Destounis S. V., DiNitto P., Logan-Young W., Bonaccio E., Zuley M. L., and Willison K. M. Can Computer-aided Detection with Double Reading of Screening Mammograms Help Decrease the False-Negative Rate? Initial Experience. *Radiology*, 232(2):578–584, 2004.

[174] Knox M., O'Brien A., Szabó E., Smith C. S., Fenlon H. M., McNicholas M. M., and Flanagan F. L. Impact of full field digital mammography on the classification and mammographic characteristics of interval breast cancers. *European Journal of Radiology*, 84(6):1056–1061, jun 2015.

[175] Weber R. J. P., van Bommel R. M. G., Louwman M. W., Nederend J., Voogd A. C., Jansen F. H., Tjan-Heijnen V. C. G., and Duijm L. E. M. Characteristics and prognosis of interval cancers after biennial screen-film or full-field digital screening mammography. *Breast Cancer Res Treat*, jul 2016.

[176] Dinitto P., Logan-young W., Bonaccio E., Zuley M. L., Willison K. M., and M R. T. R. Breast Imaging Can Computer-aided Detection with Double Reading of Screening Mammograms Help Decrease the False-Negative Rate ? Initial Experience 1. *Radiology*, 232(2):578–584, 2004.

[177] Bluekens A. M. J., Holland R., Karssemeijer N., Broeders M. J. M., and den Heeten G. J. Comparison of Digital Screening Mammography and Screen-Film Mammography in the Early Detection of Clinically Relevant Cancers: A Multicenter Study. *Radiology*, 265(3):707–714, 2012.

[178] Bria A., Marrocco C., Karssemeijer N., Molinara M., and Tortorella F. Deep Cascade Classifiers to Detect Clusters of Microcalcifications. In Et al. A. T., editor, *Breast Imaging*, volume 9699 of *Lecture Notes in Computer Science*, pages 415–422. Springer International Publishing Switzerland, 2016.

[179] Brierley J. D., Gospodarowicz M. K., and Wittekind C. *TNM classification of malignant tumours*. John Wiley & Sons, 2016.

[180] Veldkamp W. J., Karssemeijer N., and Hendriks J. H. Experiments with radiologists and a fully automated method for characterization of microcalcification clusters. *International Congress Series*, 1230:586–592, 2001.

[181] Rana R. S., Jiang Y., Schmidt R. A., Nishikawa R. M., and Liu B. Independent Evaluation of Computer Classification of Malignant and Benign Calcifications in Full-Field Digital Mammograms. *Academic Radiology*, 14(3):363–370, 2007.

[182] Hung W., Nguyen H., Lee W., Richard M., Thornton B., and Blinowska A. Diagnostic abilities of three CAD methods for assessing microcalcifications in mammograms and an aspect of equivocal cases decisions by radiologists. *Australasian Physical and Engineering Sciences in Medicine*, 26(3):104–109, 2003.

[183] Jiang Y., Nishikawa R. M., Schmidt R. a., Toledano a. Y., and Doi K. Potential of computer-aided diagnosis to reduce variability in radiologists' interpretations of mammograms depicting microcalcifications. *Radiology*, 220:787–794, 2001.

[184] Arikidis N., Vassiou K., Kazantzi A., Skiadopoulos S., Karahaliou A., and Costaridou L. A two-stage method for microcalcification cluster segmentation in mammography by deformable models. *Medical Physics*, 42(10):5848–5861, oct 2015.

[185] Fracheboud J., van Luijt P., Sankatsing V., Ripping R., Broeders M., Otten J., van Ineveld B., Heijnsdijk E., a.L.M. Verbeek, Holland R., den Heeten G., a.E. de Bruijn, and de Koning H.

Landelijke evaluatie van bevolkingsonderzoek naar borstkanker in Nederland 1990-2011/2012, 2014. URL `http://scholar.google.com/scholar?hl=en{&}btnG=Search{&}q=intitle:Landelijke+evaluatie+van+bevolkingsonderzoek+naar+borstkanker+in+Nederland{#}0`.

[186] International Agency for Research on Cancer. Cancer Screening in the European Union Report on the implementation of the Council Recommendation on cancer screening. Technical report, 2017. URL `https://ec.europa.eu/health/sites/health/files/major{_}chronic{_}diseases/docs/2017{_}cancerscreening{_}2ndreportimplementation{_}en.pdf`.

[187] Dromain C., Boyer B., Ferré R., Canale S., Delaloge S., and Balleyguier C. Computed-aided diagnosis (CAD) in the detection of breast cancer. *European Journal of Radiology*, 82(3):417–423, 2013.

[188] Jalalian A., Mashohor S. B. T., Mahmud H. R., Saripan M. I. B., Ramli A. R. B., and Karasfi B. Computer-aided detection/diagnosis of breast cancer in mammography and ultrasound: A review. *Clinical Imaging*, 37(3):420–426, 2013.

[189] Karssemeijer N., Otten J. D. M., Verbeek A. L. M., Groenewoud J. H., de Koning H. J., Hendriks J. H. C. L., and Holland R. Computer-aided detection versus independent double reading of masses on mammograms. *Radiology*, 227(1):192–200, 2003.

[190] Hupse R., Samulski M., Lobbes M., Den Heeten A., Imhof-Tas M. W., Beijerinck D., Pijnappel R., Boetes C., and Karssemeijer N. Standalone computer-aided detection compared to radiologists' performance for the detection of mammographic masses. *European Radiology*, 23(1):93–100, 2013.

[191] Melendez J., Sánchez C. I., Hupse R., van Ginneken B., and Karssemeijer N. Potential of a Standalone Computer-Aided Detection System for Breast Cancer Detection in Screening Mammography. In *IWDM '12: Proceedings of the 11th International Workshop on Breast Imaging*, volume 7361, pages 682–689, 2012.

[192] Kooi T., Gubern-Merida A., Mordang J.-J., Mann R., Pijnappel R., Schuur K., den Heeten A., and Karssemeijer N. *A comparison between a deep convolutional neural network and radiologists for classifying regions of interest in mammography*, volume 9699. 2016. ISBN 9783319415451.

[193] Helderman-van den Enden A. T. J. M., Straathof C. S. M., Aartsma-Rus A., den Dunnen J. T., Verbist B. M., Bakker E., Verschuuren J. J. G. M., and Ginjaar H. B. Becker muscular dystrophy patients with deletions around exon 51; a promising outlook for exon skipping therapy in Duchenne patients. *Neuromuscular Disorders*, 20(4):251–254, 2010.

[194] Becker A., Marcon M., Ghafoor S., Wurnig M., Frauenfelder T., and Boss A. Deep learning in mammography diagnostic accuracy of a multipurpose image analysis software in the detection of breast cancer. *Investigative Radiology*, 52(7), 2017.

[195] Kooi T., Gubern-Merida A., Mordang J. J., Mann R., Pijnappel R., Schuur K., den Heeten A., and Karssemeijer N. A comparison between a deep convolutional neural network and radiologists for classifying regions of interest in mammography. In Et al. A. T., editor, *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 9699 of *Lecture Notes in Computer Science*, pages 51–56. Springer International Publishing Switzerland, 2016. ISBN 9783319415451.

[196] Scott H. J. and Gale A. G. Breast screening: PERFORMS identifies key mammographic training

needs. *British Journal of Radiology*, 79(SPEC. ISS. 2):S127–S133, dec 2006.

[197] Cook A. J., Elmore J. G., Zhu W., Jackson S. L., Carney P. A., Flowers C., Onega T., Geller B., Rosenberg R. D., and Miglioretti D. L. Mammographic Interpretation: Radiologists' Ability to Accurately Estimate Their Performance and Compare It With That of Their Peers. *American Journal of Roentgenology*, 199(3):695–702, sep 2012.

[198] Geller B. M., Ichikawa L., Miglioretti D. L., and Eastman D. Web-Based Mammography Audit Feedback. *American Journal of Roentgenology*, 198(6):W562–W567, jun 2012.

[199] Ciatto S., Ambrogetti D., Morrone D., and Del Turco M. R. Analysis of the results of a proficiency test in screening mammography at the CSPO of Florence: review of 705 tests. *La radiologia medica*, 111(6):797–803, 2006.

[200] Reed W. M., Lee W. B., Cawson J. N., and Brennan P. C. Malignancy detection in digital mammograms. Important reader characteristics and required case numbers. *Academic Radiology*, 17 (11):1409–1413, 2010.

[201] Timmers J. M. H., Verbeek A. L. M., Pijnappel R. M., Broeders M. J. M., and Den Heeten G. J. Experiences with a self-test for Dutch breast screening radiologists: Lessons learnt. *European Radiology*, 24(2):294–304, 2014.

[202] Mordang J. J., Janssen T., Bria A., Kooi T., Gubern-Mérida A., and Karssemeijer N. Automatic microcalcification detection in multi-vendor mammography using convolutional neural networks. In Et al. A. T., editor, *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 9699 of *Lecture Notes in Computer Science*, pages 35–42. Springer International Publishing Switzerland, jan 2016. ISBN 9783319415451.

[203] Timp S., van Engeland S., and Karssemeijer N. A regional registration method to find corresponding mass lesions in temporal mammogram pairs. *Medical physics*, 32:2629–2638, 2005.

[204] Kallergi M. Using BIRADS categories in ROC experiments. In *Proceedings of SPIE*, volume 4686, pages 60–67, 2002.

[205] Varela C., Karssemeijer N., Hendriks J. H. C. L., and Holland R. Use of prior mammograms in the classification of benign and malignant masses. *European Journal of Radiology*, 56(2):248–255, 2005.

[206] Roelofs A. A. J., Karssemeijer N., Wedekind N., Beck C., van Woudenberg S., Snoeren P. R., Hendriks J. H. C. L., Rosselli del Turco M., Bjurstam N., Junkermann H., Beijerinck D., Séradour B., and Evertsz C. J. G. Importance of comparison of current and prior mammograms in breast cancer screening. *Radiology*, 242(1):70–77, 2007.

[207] Good W., Zheng B., Chang Y., and Wang X. Multi-image CAD employing features derived from ipsilateral mammographic views. In *Proceedings of SPIE*, volume 3661, pages 474–485, 1999.

[208] Paquerault S., Petrick N., Chan H.-P., Sahiner B., and Helvie M. A. Improvement of computerized mass detection on mammograms: fusion of two-view information. *Medical physics*, 29(2): 238–247, 2002.

[209] Samulski M. and Karssemeijer N. Optimizing case-based detection performance in a multiview CAD system for mammography. *IEEE Transactions on Medical Imaging*, 30(4):1001–1009, 2011.

[210] Kooi T. and Karssemeijer N. Invariant features for discriminating cysts from solid lesions in mammography. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 8539 LNCS, pages 573–580. Springer, 2014. ISBN 9783319078861.

[211] Timp S. and Karssemeijer N. Interval change analysis to improve computer aided detection in mammography. *Medical Image Analysis*, 10(1):82–95, 2006.

[212] Timp S., Varela C., and Karssemeijer N. Temporal change analysis for characterization of mass lesions in mammography. *IEEE Transactions on Medical Imaging*, 26(7):945–953, 2007.

[213] Hadjiiski L., Sahiner B., Chan H. P., Petrick N., Helvie M. a., and Gurcan M. Analysis of temporal changes of mammographic features: computer-aided classification of malignant and benign breast masses. *Medical physics*, 28(11):2309–2317, 2001.

[214] Zheng B., Tan J., Ganott M. A., Chough D. M., and Gur D. Matching Breast Masses Depicted on Different Views. A Comparison of Three Methods. *Academic Radiology*, 16(11):1338–1347, 2009.

[215] Lau T. K. and Bischof W. F. Automated detection of breast tumors using the asymmetry approach. *Computers and biomedical research, an international journal*, 24(3):273–95, 1991.

[216] Wu Y. T., Wei J., Hadjiiski L. M., Sahiner B., Zhou C., Ge J., Shi J., Zhang Y., and Chan H. P. Bilateral analysis based false positive reduction for computer-aided mass detection. *Medical physics*, 34(8):3334–3344, 2007.

[217] Wang L., Filippatos K., Friman O., and Hahn H. Fully automated segmentation of the pectoralis muscle boundary in breast MR images. In *SPIE medical imaging*, volume 7963, pages 796309–796309–8, 2011. ISBN 9780819485052.

[218] Carneiro G., Nascimento J. C., and Freitas A. The segmentation of the left ventricle of the heart from ultrasound data using deep learning architectures and derivative-based search methods. *IEEE Transactions on Image Processing*, 21(3):968–982, 2012.

[219] Pusic M. V., Andrews J. S., Kessler D. O., Teng D. C., Pecaric M. R., Ruzal-Shapiro C., and Boutis K. Prevalence of abnormal cases in an image bank affects the learning of radiograph interpretation. *Medical Education*, 46(3):289–298, mar 2012.

[220] Litjens G., Kooi T., Bejnordi B. E., Setio A. A. A., Ciompi F., Ghafoorian M., van der Laak J. A., van Ginneken B., and Sánchez C. I. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, dec 2017.

[221] Wang J., Yang X., Cai H., Tan W., Jin C., and Li L. Discrimination of Breast Cancer with Microcalcifications on Mammography by Deep Learning. *Scientific Reports*, 6(1):27327, jul 2016.

[222] Wang J., Nishikawa R. M., and Yang Y. Global detection approach for clustered microcalcifications in mammograms using a deep learning network. *Journal of Medical Imaging*, 4(2):024501, apr 2017.

[223] Geras K. J., Wolfson S., Shen Y., Kim S. G., Moy L., and Cho K. High-Resolution Breast Cancer Screening with Multi-View Deep Convolutional Neural Networks. mar 2017.

[224] Lotter W., Sorensen G., and Cox D. A Multi-scale CNN and Curriculum Learning Strategy for Mammogram Classification. pages 169–177. Springer, Cham, sep 2017.

[225] Wang J. and Yang Y. A context-sensitive deep learning approach for microcalcification detection

in mammograms. *Pattern Recognition*, 78:12–22, jun 2018.

Dankwoord

Uiteraard heb ik al die jaren niet in volledig isolement geleefd en er zijn veel mensen bij dit werk betrokken. Zonder jullie had ik dit werk nooit af kunnen krijgen en daar ben ik jullie erg dankbaar voor. De volgende personen wil ik in het bijzonder bedanken:

Prof. Dr. Karssemeijer, beste Nico, heel erg bedankt dat je mij de mogelijkheid hebt gegeven om mijn promotieambities te verwezenlijken. Bedankt voor alle begeleiding die dit boekje tot stand heeft laten komen. Ondanks dat mijn vorige promotieproject helaas niet goed liep, heb jij mij alsnog geadopteerd in jouw Breast Imaging groep. Een project dat ik erg interessant vind en vakgebied waar ik (nog steeds) heel erg veel van leer. Een pragmatischere man heb ik tot op heden nog niet mogen ontmoeten en van deze instelling heb ik erg veel geleerd. Je gaf me alle vrijheid in het onderzoek dat ik wilde doen zolang het maar met calcificaties en mammogrammen te maken had en je hebt me bijgestuurd op de momenten waar nodig. Bedankt dat ik mijn interesses in breast imaging ook verder door mag zetten binnen de industrie doordat ik een rol binnen ScreenPoint mag vervullen.

Prof. Dr. Den Heeten, beste Ard, bedankt voor je klinische visie tijdens mijn onderzoek. Dit gaf me een erg goed gevoel over de relevantie van mijn onderzoek voor de kliniek en het was een goed contragewicht om me niet in de technische details te verliezen, maar het algehele doel voor ogen te houden, namelijk het verbeteren van de borstkankerscreening.

Dr. Broeders, beste Mireille, naast de technische visie van Nico en de klinische visie van Ard, was ook jouw epidemiologische visie erg belangrijk. Ik wil je erg bedanken voor je hulp bij het plaatsen van mijn onderzoek in een groter perspectief.

Graag wil ik ook mijn manuscriptcommissie, Prof. Dr. Marchiori, Prof. Dr. Lucas en Dr. Lobbes, bedanken voor de tijd en moeite die ze hebben gestoken in het lezen en bediscussiëren van mijn thesis.

Prof. Dr. Van Ginneken, beste Bram, met jou had ik mijn allereerste gesprekken binnen DIAG en jij was in het begin van mijn DIAG-tijd één van mijn begeleiders. Jouw kritische blik en directe manier van feedback geven heeft mij voornamelijk geleerd om zelf eerst goed na te denken voordat je het met anderen bespreekt. Bedankt voor deze wijze (en soms wel harde) lessen. Samen met Nico creëerde je een goede onderzoeksomgeving voor de promovendi. Daarnaast wil ik ook graag Rashindra bedanken voor zijn enthousiasme over het onderzoek binnen DIAG tijdens mijn sollicitatie. Dit enthousiasme heeft mij uiteindelijk overtuigd om promotieonderzoek te gaan doen bij DIAG.

Prof. Torterella, dear Francesco, I would like to thank you for having me over at the University of Cassino and Southern Lazio for four months. I really enjoyed doing a project in your research group and I've learned a lot, not only about my

research topic but also about the Italian culture and cuisine. Furthermore I would like to thank Alessandro. It was very nice to have you as a partner in improving the calcification detection system. I really enjoyed working with you and I've learned a lot from you. I will forever be thankful for all your time and effort in making me feel at home (literally in your own home) in Cassino. You'll always be my Italian brother. During my stay in Cassino, I've met several people that I would also like to thank: Biaggio, Claudio, Mario, and Agnese. Thanks for making my stay in Italy unforgettable.

Uiteraard wil ik graag alle mensen bij DIAG bedanken. Met name de chille DIAG people: Nol[1], Rickert[2] , DTM[3], Stevie[4], Jan, Rientje[5] en Suuuuuuuuus[6]. De ochtendkoffie, de (tijdelijke) secret baking society en vrijdag-frietdagen zijn uiterst memorabel. Rientje, ondanks dat je DIAG vervroegd hebt verlaten, zijn we elkaar gelukkig niet uit het oog verloren. Jij gaf me een warm welkom bij DIAG en onze kanotocht blijft onvergetelijk. Rickert, bedankt voor de gezelligheid op onze kamer, het uitwisselen van de juiste werkmuziek, het uitzoeken van de beste koptelefoon en de super leuke roadtrip in Florida samen met Leti[7], Stevie en Bertje[8]. Suuuuuuuuus, ik ben het even nagegaan en totaal hebben we ieder over de jaren meer dan 47 kapsalons op, bijna 60.000 kilocalorieën de neus. Gelukkig at jij wel altijd de groente op. Verder mag ik Leti, Mohsen, Babby[9], professor Tan[10], Kaman, Thijs en Pragnya niet vergeten in dit dankwoord. Bedankt voor jullie gezelligheid tijdens mijn promotie.

Steven, bedankt dat ik je paranimf mocht zijn.

Bertje, ik denk dat je nu wel een aardig woordje Nederlands spreekt. Ik wil jou uitzonderlijk bedanken voor je hulp tijdens mijn promotie. Je hebt uit jezelf aangeboden om me te begeleiden gedurende mijn promotie en dat heeft me erg geholpen om het af te krijgen. Ik ben blij dat we nu ook weer opnieuw collega's zijn bij ScreenPoint en ik waardeer onze vriendschap die is ontstaan.

Graag wil ik alle co-auteurs bedanken die ik nog niet genoemd heb: Ritse, Jakob en Tim. Bedankt voor jullie waardevolle bijdrage aan mijn publicaties.

Colin, bedankt dat ik je al die jaren naar Nijmegen mocht chauffeuren en dat je zo nu en dan ook eens wilde rijden. Het carpoolen met jou is altijd een genot geweest,

[1]Colin Jacobs
[2]Rick Philipsen
[3]Drussen Teller Mark van Grinsven
[4]Steven Schalekamp
[5]Rieneke van Boxel - van den Boom
[6]Suzan Vreemann
[7]Leticia Gallardo Estrella
[8]Albert Gubern-Mérida
[9]Babak Ehteshami
[10]Tao Tan

niet alleen maar omdat je de mooiste man van DIAG bent, maar ook omdat het een super gezellige tijd in de auto was. Bedankt dat je mijn paranimf wilt zijn. Ik vind het fijn dat, ook al werk ik niet meer op het Radboudumc, we nog steeds zoveel mogelijk proberen te carpoolen wanneer dat kan. Ik heb een ware vriend in je gevonden en ben blij dat onze vriendschap buiten DIAG nog sterk aanwezig is.

Ter ontspanning om even het dagelijkse mentale PhD-leed te ontvluchten heb ik erg veel genoten van onze etentjes en weekendje weg met de TU/e vriendengroep. Siem[11], Rik, Channietal[12], Max, MayMay[13] en Thomas[14] bedankt voor alle gezelligheid. Ik hoop dat we deze weekendjes nog vaak mogen organiseren.

Thomas, je bent de beste vriend die ik me maar kan voorstellen. Al sinds onze studententijd zijn we vrienden en het voelt altijd oud en vertrouwd als we weer samen zijn. Onze liefde voor whisky's, speciaalbier en series kijken matched perfect. Ik kijk er dagelijks naar uit om je weer te zien. Ik voel me vereerd om je naast me te hebben als paranimf tijdens mijn verdediging. I love you, man.

Een andere vriendengroep om de ontspanning op te zoeken zijn de *de Boschenaren en de Boeman*: Peer[15], Niekers[16], Bart, Suzanne, FancyPansy[17], Lisa, VerhalliGalli[18], Jacob / JanJaap / Jozef / Mr. Bushcraft[19] en Loesje[20]. Ik hoop dat we nog vele spelletjes- en pokeravonden, escaperooms, boottochtjes en nog veel meer dagjes uit blijven plannen. Bedankt voor de fijne omgeving die jullie creëren die aanvoelt als een echte familie. Ik mag jullie wel!

Uiteraard wil ik ook mijn echte, echte familie bedanken. Mijn broers en zus, Martin, Richard en Nikki, en mijn ouders. Lieve pap[21] en mam[22] bedankt voor al jullie zorg en levenslessen die ervoor gezorgd hebben dat ik ben zoals ik nu ben. Jullie hebben me alles gegeven dat ik nodig heb om me staande te kunnen houden tegen wat dan ook. Graag wil ik ook Sandra, (tante) Karin en (tante) Lilian bedanken voor hun steun.

Als laatste wil ik natuurlijk Mies[23] bedanken. Pfoe, wat een lange rit was die promotie hè? Ik wil je ontzettend bedanken voor alle steun de je me hebt gegeven tijdens

[11] Simone Booij- Bouwmans
[12] Chantal Tax
[13] May Wong
[14] Tom Dela Haije
[15] Peter den Bieman
[16] Daniek Richt
[17] Fanny le Blanc
[18] Michel Verhallen
[19] Doorhalen welke op dit moment niet van toepassing zijn
[20] Marloes van Dun
[21] John Mordang
[22] Trees Mordang - Rijken
[23] Micheline ten Thij (en nu: Micheline Mordang)

mijn promotieonderzoek. Ik zit nu in een vrij unieke situatie waarbij ik je op het moment van schrijven mijn verloofde mag noemen en op het moment van publiceren mijn vrouw mag noemen! Ik hoop dat we nog oneindig veel Mies & Jurre-avonturen gaan beleven in de toekomst. Ik hou van je, lief Miepelientje.

Curriculum Vitae

Jan-Jurre Mordang was born in 's-Hertogenbosch, the Netherlands, on April 8th 1988. In 2006 he started the study Biomedical Engineering at the Eindhoven University of Technology (TU/e). After his bachelor degree, he started the Medical Engineering master and his master thesis entitled *"Evaluation of non-invasive methods for quantification of skeletal muscle metabolism and oxygenation in patients with chronic heart failure"* in collaboration with the TU/e and MMC Veldhoven. In September 2011, he started as a PhD student in the Diagnostic Image Analysis Group in the Department of Radiology at the Radboud University Medical Center. His first project was about 4D CT of the brain. In March 2013, he switched to computer aided detection in mammography where he mainly worked on the automatic detection of calcifications. During his PhD program, he has visited the University of Cassino and Southern Lazio for a four month project which resulted in a chapter of this thesis. After his PhD, Jan-Jurre started working at ScreenPoint Medical, a company that develops deep learning and image analysis technology for automated reading of mammograms and digital breast tomosynthesis.