Computer Aided Detection as a Decision Aid in Medical Screening

Maurice Samulski

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

Cover design: Anne-Gudrun Klæth Lyngsmo
Book layout: Maurice Samulski

ISBN-10: 82–497–0314–6
ISBN-13: 978–82–497–0314–2

Printed by Ipskamp Drukkers, Nijmegen.

COMPUTER AIDED DETECTION AS A DECISION AID IN MEDICAL SCREENING

Een wetenschappelijke proeve op het gebied van de
Natuurwetenschappen, Wiskunde en Informatica

PROEFSCHRIFT

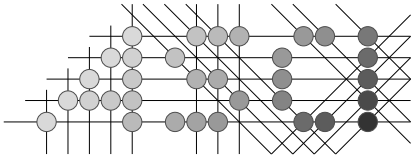
ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus, prof. mr. S.C.J.J. Kortmann,
volgens besluit van het college van decanen
in het openbaar te verdedigen op woensdag 14 december 2011
om 10.30 uur precies

door

MAURICE RENÉ MARINA SAMULSKI

geboren op 26 december 1978
te Kerkrade

Promotoren:	Prof. dr. ir. N. Karssemeijer
	Prof. dr. P.J.F. Lucas, MD (Leiden University)
Manuscriptcommissie:	Prof. dr. T.M. Heskes
	Prof. dr. ir. B.P.F. Lelieveldt (Leiden University Medical Center)
	Prof. dr. M. Prokop, MD



The research described in this thesis was carried out at the Diagnostic Image Analysis Group, Radboud University Nijmegen Medical Center (The Netherlands), and the Advanced School for Computing and Imaging (ASCI) graduate school. ASCI dissertation series number 241.

This work was funded by grant KUN 2006-3655 of the Dutch Cancer Society sponsored by fundraising activities of cycling club "Bergh in het Zadel" and the Dutch Organization for Scientific Research (NWO) under BRICK/FOCUS grant number 642.066.605.

Financial support for publication of this thesis was kindly provided by the department of radiology of the Radboud University Nijmegen Medical Centre.

Contents

Introduction

1

1.1 Medical imaging

There is an exponential growth of medical image data being produced in current clinical practice for effective patient diagnosis. The clinical benefit of medical image data for patient care is largely dependent on the quality of the images acquired and the ability of the radiologist to interpret them. For many years, it has been recognized that even the best radiologists make errors in the interpretation of medical exams. The errors that are being made include perception failures and interpretation failures[1–5]. Perception errors occur when an abnormality is in the field of view of the reader but remains undetected. Interpretation errors occur when an abnormality is detected, but is incorrectly interpreted as normal or benign. These failures can be attributed to radiologist fatigue, limitations in the human visual system, distractions, experience level, overlapping glandular tissue that (partly) obscures abnormalities, and large volumes of normal cases in a screening situation.[6–12] To reduce these problems, computer aided detection and diagnosis systems have been designed to help improve detection performance. In the following sections I will describe in detail the history and current developments on computer aided detection (CADe) and diagnosis (CADx) methods.

1.2 History of computer aided detection

Early attempts at analyzing of abnormalities in radiographic exams with computers were made in the 1960s.[13] In the 1960s and 1970s computers were used to automatically detect and classify abnormalities in medical images, including mammograms and chest radiographs.[14–17] It was thought that computers could replace radiologists in detecting abnormalities. However, these early attempts were not so successful. The inferior quality of medical images and limited computing power may have had a detrimental effect on the success of these early attempts.

In the 1980s, computerized methods were developed with the intention to aid the radiologists rather than completely automatic interpretation by computers, focussing initially on detecting lesions in mammograms and chest radiographs.[18,19] In this form, the radiologists used the output of the CAD system mainly to avoid overlooking abnormalities. It is important to note that the final decision was made by the radiologist, and not the computer system, *i.e.* CAD was not intended to replace the radiologist, but was used to aid the radiologist in the detection and interpretation of abnormalities.

A large number of studies have been published in the field of CAD over the past 20 years, going from coarse laboratory tools evaluated on a small number of cases, to sophisticated CAD schemes and commercial CAD systems that are evaluated on large clinically relevant databases.

1.3 CAD techniques

Broadly speaking, there are two types of CAD techniques: computer aided detection (CADe) and computer aided diagnosis (CADx). CADe systems have been developed to aid radiologists in localizing suspect regions, leaving the characterization and diagnosis to the radiologist. It only gives the location of suspect lesions by showing prompts to the radiologist, and serves as an aid in the *detection* task. CADe is most beneficial in screening situations in which many cases need to be interpreted by radiologists, but where many cases are normal, such as breast cancer screening, colon cancer screening, and lung cancer screening.

CADx systems are used as an aid to further interpret a region or lesion that is already located by either the radiologist or a CADe system. CADx serves as an aid in the *classification* task for differential diagnosis. These systems often present the radiologist with an estimated probability of malignancy of a suspect region, classifying it into possibly malignant or likely benign. In addition there are other ways of presenting the CAD results, which will be discussed in section 1.4.6.

The overall goal of CAD is to reduce the number of perception errors, reduce the number of interpretation errors, and to reduce the variability between radiologists.

1.4 CAD in breast imaging

CAD has been in use in mammography since the beginning of the CAD research era. The interpretation of mammograms is one of the most difficult tasks in radiology. Abnormal mammographic signs such as calcifications and masses can be very subtle and are often obscured by dense fibroglandular breast tissue. In the Netherlands, asymptomatic women aged 50 to 75 years are invited for a mammographic examination of both breasts on a bi-annual basis. Most of the screening mammograms are taken in mobile mammography units throughout the Netherlands, but some are performed at hospitals. These include the high risk group of women with a genetic predisposition of breast cancer, or a family history of breast cancer. Mammograms can be taken from different angles, where the most common are the mediolateral oblique (MLO) view and the craniocaudal (CC) view. The MLO view is taken from the side at an angle between 30 and 60 degrees, and shows more of the breast tissue to be imaged. Also part of the pectoral muscle is visible in the MLO view. The CC view is taken from above, and sometimes the area close to the chest wall is not visible. Breast cancer screening programs have been surrounded by long-running debates concerning its benefit and harm.[20,21] The large volume of cases that need to be interpreted in current screening makes missing subtle signs of breast cancer a real possibility, and it is thought that

CAD could reduce the chance that radiologists will overlook cancers.

In retrospective reviews in which readers were blinded and read screening mammograms from women who subsequently developed breast cancer mixed with normal mammograms it was shown that radiologists missed an estimated 20-32% of cancers that should have been recalled.[22–25] The study by Brem *et al.*[24] indicated that 32% of the prior mammograms (123/377) had actionable findings. A study by Warren-Burhenne[22] showed similar results, in which 27% (115/427) of screen detected cancers that had findings visible in the prior mammogram were deemed actionable. Retrospective nonblinded reviews show that up to 75% of the missed cancers were visible on the prior screening mammogram.[8,26–29] CADe, therefore, could potentially be very useful in mammography to reduce the number of false-negatives.

Most CADe systems in mammography focus on the detection of either clustered micro-calcifications or masses. In general, it is accepted that mass detection is a more challenging problem than the detection of micro-calcifications, because of the large variation in appearance of the masses in mammograms and their low contrast compared to their surroundings.[30,31] In recent years there are also efforts to specifically design CADe schemes for detecting architectural distortion.[32–35] These efforts have been described extensively in many publications, and the review paper from Giger *et al.* gives a complete overview of these efforts.[36]

CAD for the detection of micro-calcifications has been extensively investigated by a number of groups, and vary in specific techniques used.[37–39] In general, one can identify some common steps. First the mammogram is segmented into breast tissue and the background area. Then often the noise is suppressed to enhance the contrast between the micro-calcifications and background. Candidate selection is performed, and the micro-calcifications are segmented to extract features that characterize the shape, size, contrast, and number of micro-calcifications in a cluster. Additional false-positive reduction techniques are subsequently employed such as neural network classifiers to distinguish between suspicious micro-calcifications and normal tissue.

CAD for the detection of masses has received a lot of attention since beginning 1990s.[40–47] In general, a mass detection scheme contains several steps that are similar to the micro-calcification detection task. As a first step, the mammogram is segmented into breast tissue, pectoral muscle (if the image is a MLO view), and background area. Commonly a pixel classifier is trained with a small set of features, and each pixel in the breast tissue is classified resulting in a likelihood map. Using a threshold, candidate locations are selected from this likelihood map and segmented. For each of the candidate objects, various features describing the border, contrast, location, texture, presence of spiculation are calculated. Classifiers, such as the neural network, are used to classify the mass candidates as a true-positive or false-positive. Many CADe schemes

described in literature are based on two stage classification approaches.

Most of the current CAD systems detect suspicious lesions independently in single views. However, in clinical practice radiologists combine information from all available views. They compare MLO and CC projection views, look for asymmetries in bilateral mammograms, and compare the current mammograms with the prior mammograms to identify changes. Using all views improves the chance of detecting abnormalities and reduction of false-positives. There are several studies conducted that try to emulate the radiologist' practice, and incorporate information from multiple views (bilateral[42,48–50] and ipsilateral[51–54]) to improve detection performance. Comparing current with prior mammograms is done routinely by radiologists, to detect new or growing abnormalities, and consequently there have been efforts to incorporate temporal information into CAD schemes.[55–57]

1.4.1 Observer studies

The first observer study that investigated the effect of CAD on the radiologists' performance was the study in 1990 from Chan *et al.*[58] in which the performance of the radiologists on the detection of micro-calcifications was compared with and without the aid of the CADe system. They demonstrated that the detection performance significantly increased with the use of CADe, and showed the potential usefulness of CADe systems as a second opinion. It was also clear from this study that it was not necessary for the CADe system performance to be as high or higher than that of the radiologists to be an useful aid, as long as it can provide complimentary information. In 1994, a similar observer study was performed by Kegelmeyer *et al.*[59] that investigated the influence of CAD for the detection of masses, and since then more studies followed that showed the positive effect on the detection performance of radiologists.[60–62]

1.4.2 Potential of CAD

The potential of CAD to identify malignant masses that have been missed by radiologists was presented by Schmidt in 1996 *et al.*.[63] Retrospective analysis showed that 54% of the 69 missed cases in clinical practice were identified by CAD. In a similar study from 1998, te Brake and colleagues[64] showed that 34% of the 65 cancers that were initially missed by two radiologists were detected by their research CAD system. Warren-Burhenne and colleagues[22] reported that a more recent version of CAD (ImageChecker M1000, R2 Technology, Los Altos, CA) successfully marked the missed cancers in 77% of the false-negative prior mammograms. In the study by Brem *et al.*[24], the computer-aided detection system detected 63% of the 123 missed cancer cases that

had a subsequent screening mammogram that led to a cancer diagnosis. These results suggested that 80 of the 123 cancer cases could have been identified by the radiologists if they had used CADe.

1.4.3 Prospective clinical trials

The first commercial CADe system for screening mammography was approved by the Food and Drug Administration (FDA) in 1998. After that time, also other systems for mammography have obtained FDA approval[36]. While the use of CADe systems for screening mammograms has been steadily increasing in the clinic, there is ongoing controversy regarding its clinical benefit. Several studies showed that CADe has a strong benefit in a laboratory setting, but there are mixed reports on the actual clinical performance.[65] Several reports have been published on the performance of these commercial systems in clinical practice.

The CADET II study[66] was a large multi center prospective randomized trial in which single reading with CAD was compared with double reading involving 31 057 mammograms. This study design reflects more closely the effects of CAD when it is introduced into screening mammography. The results of this study showed that there was no significant difference in cancer detection rate between single reading with CAD and double reading, and that the recall rate for single reading with CAD was 0.5% higher than that for double reading. This shows that single reading with CAD could be an alternative to double reading. In an other large prospective study by Gromet[67] (231 221 mammograms) CADe increased the sensitivity of a single reader, with only a small increase in recall rate. In a prospective study in the UK involving 18 096 cases, Khoo *et al.*[68] showed that single reading with CAD increased the sensitivity over that of single reading alone by 1.3%. However, double reading by radiologists increased sensitivity by 8.2%. Morton *et al.*[69] determined prospectively that the use of CADe increased breast cancer detection by 7.62%, and increased the recall rate from 9.84% to 10.77% (relative increase of 9.5%). These studies used a sequential reading design where cases were read first without CAD, immediately followed by an interpretation where CADe results were displayed. Therefore, the results were collected from the same patients and the same radiologists.

In studies where the detection performance of radiologists were compared over two periods of time, before and after CADe was implemented in practice, the results are less optimistic. A study by Gur *et al.*[70], in which 24 radiologists interpreted 59 139 screening mammograms with, and 56 432 without CADe, reported no statistically significant change in breast cancer detection rate. In this study design different patients are screened, and the radiologists may not be the same in both periods. A more re-

cent study showed even a detrimental effect on radiologists' performance when CADe was used. Fenton *et al.*[71] found that the implementation of CADe was associated with significant increase in recall rate (relative increase of 30.7%) and biopsies with no clear impact on the early detection of breast cancer (relative increase of 4.5%). The various outcome of these large published prospective studies are causing a continuous debate regarding the usefulness of CAD in its current form.

1.4.4 Computer aided diagnosis

Computer-aided diagnosis (CADx) methods are used to aid the radiologist in characterizing an already found abnormality, and in the estimation of its probability of malignancy. Using this additional information radiologists can make a differential diagnosis (*e.g.* if a biopsy is justified to further assess the abnormality). The input of a CADx algorithm could be either a radiologist-detected or a computer-detected location, from which lesion features are extracted. These features are fed into a classifier and result into a diagnosis. Already in 1988, Getty *et al.* showed that when lesion characteristics that were given by radiologists were merged by a classifier the performance of the radiologists improved.[72] In current CADx systems, often computer-extracted features are used that are calculated from a radiologists' delineated region or an automatically segmented region from *e.g.* the CADe system[73–76]. Examples of these features are presence of spiculation, micro-calcification cluster distribution, contrast differences, presence of texture, border characteristics, and others.[77]

1.4.5 Observer studies

As with computer-aided detection, there are various observer studies conducted investigating the effectiveness of CADx as an aid to radiologists in the task of distinguishing between malignant and benign lesions.[61,62,78,79] In all these studies, it was found that the use of CADx significantly improved the radiologists' performance using mammography or breast ultrasound. Recently there are also observer studies published that investigated the potential of multi-modal CADx, combining the diagnosis of mammography and breast ultrasound systems.[80,81] Multiple CADx output can be given separately for these different modalities or be combined into one score using the features of the available modalities.[82]

1.4.6 Presentation of CAD results

Conventional CADe systems are used as a technology to help radiologists to avoid perceptual errors by providing them with visual CADe marks after they have initially

evaluated the case and a preliminary mental decision is made whether a case should be recalled or not. The assumption is that significant lesions missed by radiologists, will be acted upon when CADe marks them. A major complaint from radiologists is that current CADe systems generate a lot of false-positive prompts, and often cases are referred based on false CADe prompts while rejecting true CADe prompts. In addition, many lesions are not missed by perceptual oversight but due to misinterpretation[5,83,84]. Therefore, conventional CADe may not be the most effective way to avoid missing cancers, and the current concept may need to be revised to prevent interpretation errors in the screening.[85] One of the goals of this thesis is to design an interactive computer-aided detection system that helps the radiologist with decision making, and to investigate the potential benefits of this approach on the detection performance of radiologists. Instead of using the traditional prompting approach[86,87], where CAD results are displayed after the radiologist has evaluated the case, in the interactive approach CAD results are only displayed on request during reading. Basically, this means that the radiologist reviews the mammogram and selects areas that attracts his attention to scrutinize further with the help of the CADe system. When a certain area in a mammogram is queried, CAD information about this location will be displayed if available. If available, the contour of the region is shown with a level of suspicion that is computed by the CAD system. This interactive approach is mainly addressed to avoid interpretation errors, and obviously does not avoid perceptual oversights.[88] A major advantage of this approach is that the radiologists are not burdened with many of the false-positives of CAD.

The way of presenting CADx results to the radiologist is important allowing the radiologist to make an optimal decision. In literature there are CADx systems described that present the radiologist with an estimated probability of malignancy of a suspect region classifying it into possibly malignant or likely benign, systems that retrieve masses that are similar to the inspected region from a reference database based on computer extracted features[89–92], give qualitative information about features that describe the suspect region, give a representation of the case in question relative to the distribution of the normal and abnormal cases in a given population, or a combination of those[80,93].

1.5 CAD in lung imaging

Chest radiography is the most common imaging technique for the diagnosis of pulmonary diseases, mainly due to low cost and short examination time.[94] However, the detection of pulmonary nodules at an early stage in chest radiographs is an extremely

difficult task for radiologists, and up to 90% of the cases that were missed contained nodules that were visible in retrospect.[95–99] CAD for lung disease is an active field of research, beginning in the 1970s.[16,17,100–105] The effect of CAD for lung nodule detection on radiologists has been extensively investigated and, similar to the studies in mammography, showed that the radiologists' performance could be increased using CAD.[100,106–112] In 2001, the first commercial lung nodule CAD system was accepted by the FDA. However, no large prospective trials have been published that evaluated the performance of radiologists using this CAD system in clinical practice, to my knowledge.

It is expected that CT has a higher sensitivity for the detection of lung nodules than chest radiographs, and there have been efforts to develop CAD systems for lung nodule detection on CT scans.[113–115] The reported performances vary between studies, due to different scanning protocols and different data set compositions. Recent retrospective observer studies indicated that the radiologist performance increased significantly using CAD for lung nodules on CT scans.[116,117] There are also CAD systems for the diagnosis of lung nodules investigated, which schemes are comparable to the schemes for CADx in mammography.[118–121]

1.6 CAD in other diseases

Another important area of CAD applications is in colon imaging. CADe systems detect polyps that can be a precursor of colon cancer, one of the leading causes of cancer deaths in the world. The most reliable technique to date is the invasive technique of colonoscopy. As an alternative, CT colonography (CTC) is being investigated. However, the interpretation of CTC is time consuming and difficult. Therefore, CADe systems may be useful to aid the radiologist interpreting CTC to reduce false-negatives and reader variability. The performance of CADe systems for detecting polyps in CTC vary a lot among studies, partly due to the small size of data sets that were used to validate the developed systems. There are no large scale prospective trials reported in the literature, but several retrospective observer studies indicate the potential of CADe to improve radiologists' performance.[122,123]

1.7 CAD evaluation

In CAD research, we often have to assess the diagnostic performance of a CAD system, evaluate differences between CAD systems, and determine the performance of radiologists using a CAD system as an aid in their decisions. There are a wide range

of these performance evaluation techniques used in this thesis with abbreviations such as ROC, FROC, LROC, and JAFROC, which have been referred to as an alphabet soup by Xin He and colleagues. In this section I will briefly describe the evaluation methods that are used in this thesis.

Receiver operating characteristic (ROC) analysis is a statistical method for analyzing, visualizing and comparing the performance of binary classification tasks. To evaluate the performance of a CADx system or observer study, where the task is to classify a suspicious region into benign or malignant, ROC analysis is typically used. CADe methods are usually evaluated with free-response ROC (FROC) analysis to visualize the relation between sensitivity and the average number of false-positives per image or case, accounting for the localization and detection of abnormalities.

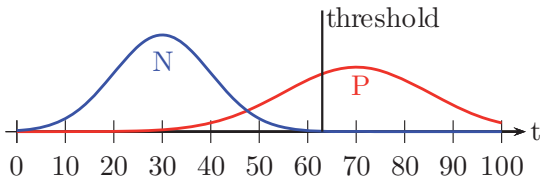


Figure 1.1: An example of two distributions of rating values for negative and positive cases.

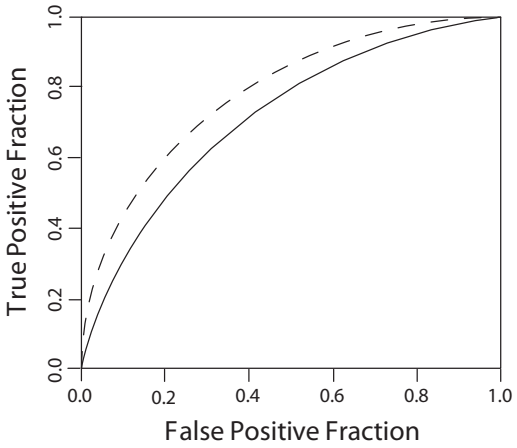


Figure 1.2: Two examples of a receiver operating characteristic (ROC) curve. The system with the dashed ROC curve performs better than the system with the solid ROC curve.

In a ROC study, a radiologist or CADx system assigns a rating to each image that represents the likelihood that the image contains an abnormality[124]. A plot can then be made that shows the rating value distributions of the two classes (benign, malignant), of which an example is shown in Figure 1.1. To construct a ROC curve, the true-positive fraction (TPF) and false-positive fraction (FPF) are computed for each possible threshold in Figure 1.1, resulting in points through which a curve is drawn or fitted. Examples of ROC curves are shown in Figure 1.2. The CADx system that produced the dashed ROC curve is superior to the one that produced the solid curve, since for every decision threshold the dashed curve is above the other curve, *i.e.* for any given specificity the sensitivity is higher. A commonly used measure to summarize the ROC curve into one number, is the area under the ROC curve (AUC or A_z value). An AUC

value of 0.5 indicates that the system does not perform better than random chance, and a system with an AUC value of 1 has perfect diagnostic performance. When ROC curves do not cross, a higher AUC value means better performance in comparison to a system with a lower AUC value. However, when ROC curves do cross the AUC only indicates the average performance, and the diagnostic task you want to perform becomes very relevant. For example in cancer screening, radiologists perform at a very low false-positive rate (*i.e.* at high specificity), making the left part of the curve more important. In Figure 1.3 an example is given of such situation. The technique producing the dashed curve is better in a screening setting than the solid curve. In such situations a partial area under the curve value can be computed to compare system performances.

In a classic ROC study, an observer assigns a rating to each image and can only be applied if the location of the abnormality in the image is known or is not important. In diagnostic tasks where location is important, an extension of ROC is available: localization ROC (LROC).[125] In a LROC study an observer is asked to mark to location of the suspicious region and provide a rating. The mark provided by the observer is considered a true-positive if the mark location is close enough to the true location (ground truth). In a LROC plot the correctly localized true-positives are plotted versus the false-positive fraction. An example is shown in Figure 1.4. As can be seen, the curve does not necessarily end at 100% sensitivity in comparison to the ROC curve, because some true-positives could not have been correctly localized. Also in LROC curves, the (partial) area under the curve could be computed to compare diagnostic performances, with the same considerations as for ROC curves when the curves cross. An alternative measure that can be used, is the mean sensitivity in a false-positive fraction interval.

To evaluate the performance of a CADe system, usually a free-response ROC (FROC) is employed to understand the relation between sensitivity and average false-positives per image.[126] In a FROC study, the observer (*e.g.* a CADe system or radiologist) can report multiple suspicious locations and give them a rating, and it is possible that there is more than 1 lesion per image as opposed to LROC analysis. To construct a FROC curve from this data, the fraction of correctly localized true-positives is computed for each decision threshold, with the associated average number of false-positives per image. A CAD system is better when the curve is higher in the vertical direction (*i.e.*, more correctly localized true-positive decisions) and steeper (*i.e.*, with fewer false-positives per image). The area under the FROC curve cannot be used to evaluate performance, as more false-positives per image also results in a higher area under the FROC curve. Alternative measures used in literature are the partial area under the FROC curve, or a mean true-positive fraction (MTPF) in a certain false-positives per image interval. Original FROC analysis is lesion-based (also sometimes referred to as image-based),

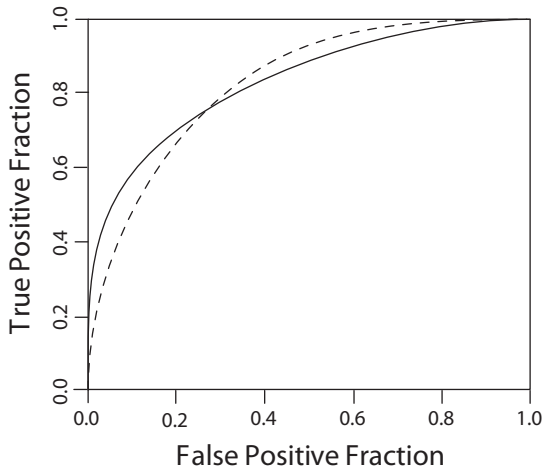


Figure 1.3: Two crossing ROC curves with approximately the same AUC value. Which curve is better is dependent on the diagnostic task at hand.

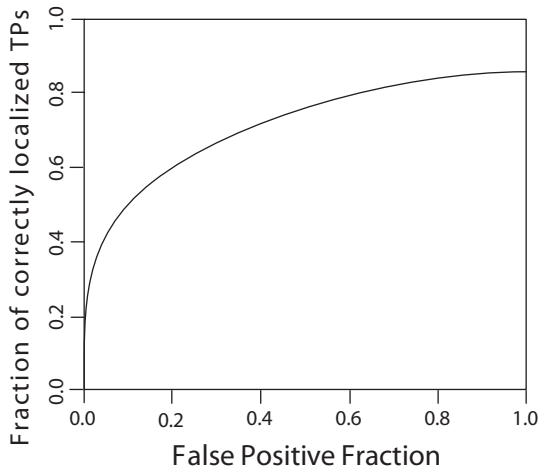


Figure 1.4: An example of a localization ROC curve, often used in observer studies.

and requires the observer to find the lesion in all available projections. More relevant in clinical practice is that at least one lesion is found in a patient. Therefore, a lot of CAD studies perform case-based FROC analysis (Figure 1.5). The only difference of case-based FROC curves in comparison to the traditional lesion-based FROC curves is that on the y-axis the fraction of cases (*i.e.* patients) is shown in which the observer correctly localized a lesion in at least one of the projection views.

Alternative FROC (AFROC) analysis is a variant of FROC analysis that has the same vertical axis, but a different definition of the horizontal axis.[127] In a FROC curve, the horizontal axis extends to an arbitrary large number of false-positives per image or case. In an AFROC curve the horizontal axis indicates the fraction of normal images containing 1 or more false-positive reports and only the false-positive with the highest rating is considered. For each decision threshold, the fraction of correctly localized true-positives is computed and the associated fraction of negative images that have been falsely recalled. Analogue to ROC studies, the area under the AFROC curve can be used for performance assessment.

1.7.1 Statistical evaluation

Many statistical tests have been developed to assess if the difference between two ROC curves is statistically significant. Some of these methods (*e.g.* a paired or unpaired Student's *t* test) take only reader variation into account, which means that the drawn conclusion can be applied to readers in general, but only to the studies' specific selection of

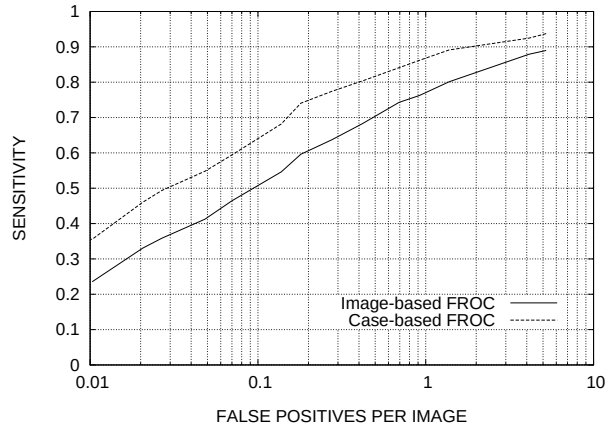


Figure 1.5: An image-based and case-based free-response receiver operating characteristic (FROC) curve are shown of an example CAD system. On the logarithmic x-axis, the number of false-positives per image are shown. On the y-axis the image sensitivity is shown for an image-based FROC, and for a case-based FROC the case sensitivity is shown.

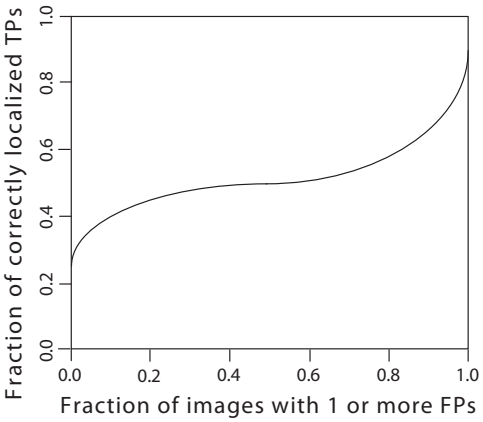


Figure 1.6: An example of an alternative FROC (AFROC) curve, which is similar to the FROC curve except for its horizontal axis.

cases. Conclusions drawn from methods that take only case variation into account can be applied to cases in general, but only to the specific readers that participated in the observer study (*e.g.* the ROCKIT software[128]). Better ROC analysis methods take both reader and case variation into account, which are often called multiple reader, multiple case (MRMC) methods. A commonly used software package is DBM MRMC[129] from the University of Chicago which uses analysis of variance (ANOVA) methods together with jackknifing to assess the statistical significance of the observed difference between two systems.

Statistical methods to evaluate differences between (alternative) free-response ROC curves are increasingly used and remain an active research topic.[130–135] A non-parametric method to obtain the p-value and confidence intervals is the well-established bootstrap method.[134–136] In this method, cases from the test set are sampled with replacement a sufficient number of times (*e.g.* 5000), and for each resampling two FROC curves for the systems under test are constructed and compared. The figure of merit that is used in this thesis is the mean true-positive fraction (MTPF) within a particular false-positive range of the FROC curve, but also other evaluation metrics can be used. After the resampling procedure, a large number of differences in performance values are available from which confidence intervals and the p-value can be derived. This non-

parametric method is preferred when CAD algorithms are being compared.[137] For human observer FROC data, an other well-known method is used: the jackknife alternative FROC (JAFROC) method[133], where the figure of merit is the area under the AFROC curve, and the statistical significance is determined using ANOVA analysis similar to the DBM MRMC method.

1.8 Outline

The main objective of this thesis was to optimize computer aided detection techniques using information from ipsilateral views and to develop and evaluate an alternative method of presenting computer aided detection results to the radiologists. In current screening, computer aided detection systems display suspicious mammographic regions as prompts to the radiologists with the intention to avoid perceptual oversights. However, it has been shown in the literature that interpretation errors might be a more common cause of missing cancers in screening. In this thesis, we investigated an alternative paradigm of a computer-aided detection system that presents CAD information for a suspicious region on request while the radiologist is reading the case. This method is aimed to help with decision making, rather than to avoid overlooking cancers.

The outline of this thesis is as follows. In chapter 2 we compared Bayesian networks with support vector machines characterizing mammographic lesions as either benign or malignant. In this chapter the components of our single view detection and classification scheme are described. In Chapter 3 we investigated whether a reliable likelihood measure for a patient being cancerous could be obtained by combining information available as detected regions from a single-view CAD system from both mammographic views using a Bayesian approach. In Chapter 4 a probabilistic approach is presented to link suspicious regions detected by a single-view CAD system in MLO and CC views based on their correspondence information. In Chapter 5 we investigated a multi-view scheme to improve case-based mass detection performance of our single-view CAD system by optimizing the selection of training patterns based on correspondence information from the previous chapter. In chapter 6 an alternative paradigm of computer-aided detection systems is proposed where radiologists are helped with decision making instead of preventing perceptual errors and its effect on the radiologists' performance in a screening setting is evaluated. In Chapter 7 the traditional paradigm of CADe (prompting) is compared to the interactive usage of CADe in mammography. Chapter 8 explores whether this new CAD concept can also improve detection performance of lung nodules in chest radiography.

Classification of mammographic masses using SVM and Bayesian networks

2

Maurice Samulski, Nico Karssemeijer, Peter Lucas, and Perry Groot

Original title: Classification of mammographic masses using support vector machines and Bayesian networks

Published in: Proceedings of SPIE Medical Imaging 2007: Computer-Aided Diagnosis. Volume 6514(1), pp. 65141J

Abstract

In this paper, we compare two state-of-the-art classification techniques characterizing masses as either benign or malignant, using a dataset consisting of 271 cases (131 benign and 140 malignant), containing both a MLO and CC view. For suspect regions in a digitized mammogram, 12 out of 81 calculated image features have been selected for investigating the classification accuracy of support vector machines (SVMs) and Bayesian networks (BNs). Additional techniques for improving their performance were included in their comparison: the Manly transformation for achieving a normal distribution of image features and principal component analysis (PCA) for reducing our high-dimensional data. The performance of the classifiers were evaluated with Receiver Operating Characteristics (ROC) analysis. The classifiers were trained and tested using a k-fold cross-validation test method (k=10). It was found that the area under the ROC curve (A_z) of the BN increased significantly (p=0.0002) using the Manly transformation, from $A_z = 0.767$ to $A_z = 0.795$. The Manly transformation did not result in a significant change for SVMs. Also the difference between SVMs and BNs using the transformed dataset was not statistically significant (p=0.78). Applying PCA resulted in an improvement in classification accuracy of the naive Bayesian classifier, from $A_z = 0.767$ to $A_z = 0.786$. The difference in classification performance between BNs and SVMs after applying PCA was small and not statistically significant (p=0.11).

2.1 Introduction

Machine learning techniques to diagnose breast cancer is a very active research area. Several computer-aided diagnosis (CAD) systems have been developed to aid radiologists in mammographic interpretation. These CAD systems analyze mammographic abnormalities and classify lesions as either benign or malignant in order to assist the radiologist in the diagnostic decision making. Some of them are based on Bayesian networks learned on mammographic descriptions provided by radiologists[138] or on features extracted by image processing[139]. Another classification technique that is widely used for the diagnosis of breast tumors are support vector machines[140–143]. The theoretical advantage of SVMs is that by choosing a specific hyperplane among the many that can separate the data in the feature space, the problem of overfitting the training data is reduced. They are often able to characterize a large training set with a small subset of the training points. Also, SVMs allow us to choose features with arbitrary distributions, and we do not need to make any independence assumptions. The advantage of Bayesian networks is that statistical dependences and independences between features are represented explicitly, which facilitates the incorporation of background knowledge. In this study we compare both classification methods and use two techniques, namely dimension reduction by principal component analysis (PCA) and a transformation for achieving a normal distribution of image features, to further improve the accuracy rate of the classifiers. Recently, the combination of PCA and support vector machines (SVMs) has been used in medical imaging, where principal component analysis is applied to extracted image features and the results are used to train a SVM classifier, but not specifically for mammograms[144].

2.2 Materials and methods

The digitized mammograms that were used in this study have been obtained from the Dutch Breast Cancer Screening Program. In this program two mammographic views of each breast were obtained in the initial screening: the medio-lateral oblique (MLO) view and a cranio caudal (CC) view. In this study 271 cases were used. Of these cases, 131 were benign and 140 were malignant. All cases had four-view mammograms.

To each image in the dataset a CAD scheme was applied that was previously developed in our group[145]. The CAD scheme consists of the following steps (Figure 2.1):

- Segmentation of the mammogram into breast tissue, pectoral muscle (if image is a MLO view), and background area

- Initial detection step resulting in a likelihood image and a number of suspect image locations (local maxima)
- Region segmentation, by dynamic programming, using the suspicious locations as seed points
- Final classification step to classify regions as true abnormalities and false positives.

These steps will be described in more detail in the following subsection.

2.2.1 Likelihood detection

Segmentation of the mammogram The first step of our CAD scheme is the segmentation of an image into breast tissue and background, using a skin line detection algorithm. Additionally, it finds the edge of the pectoralis muscle if the image is a MLO view.[146] After these steps, a thickness equalization algorithm is applied to enhance the periphery of the breast[147]. A similar algorithm is used to equalize background intensity in the pectoralis muscle, to avoid problems with detection of masses located on or near the pectoral boundary.

Initial mass detection step In this step we use a pixel-level method: for each pixel inside the breast area there are a small number of features calculated that represent presence of a central mass and the presence of spiculation[44]. A neural network classifies each pixel using these features and assigns a level of suspiciousness to it. The neural network is trained using pixels sampled inside and outside of a representative series of malignant masses. The result is an image in which pixel values represents the likelihood that a malignant mass or architectural distortion is present. This likelihood image is then slightly smoothed and a local maxima detection is performed. A local maximum is detected when the likelihood is above a certain threshold and no other nearby locations have a higher likelihood value. This results in a number of suspicious locations. Finally an algorithm searches for local maxima that are located closer than 8 mm together and remove multiple candidate locations to avoid multiple suspicious locations on the same lesion.

Region segmentation Each of the detected local maxima in the previous step are used as seed points for region segmentation, based on dynamic programming[148].

Final classification For each segmented region, 81 features are calculated related to lesion size, roughness of the boundary, linear texture, location of the region, contour

smoothness, contrast, and other image characteristics. In the conducted experiments

we used a subset of 12 features out of 81 features. They were selected using a k-nearest neighbor (KNN) algorithm and sequential forward procedure to find the most useful features for classifying lesions as benign or malignant. The procedure is described in detail in previous research[149]. We will give a short description of the used features in the following subsection.

2.2.2 Region features

Spiculation features Malignant mammographic densities are often surrounded by a radiating pattern of linear spicules. For the detection of these stellate patterns of straight lines directed toward the center pixel of a lesion, two features have been designed by Karssemeijer and te Brake[44]. The idea is that if an increase of pixels pointing to a given region is found then this region may be suspicious, especially if, viewed from that region, such an increase is found in many directions. The first feature *Stellateness 1* is a normalized measure for the fraction of pixels with a line orientation directed towards the center pixel. We call this set of pixels F. For calculating the second feature *Stellateness 2*, the circular neighborhood is divided into 24 angular sections. This feature measures to what extent the pixels in set F are uniformly distributed among all angular sections. Also the mean values of *Stellateness 1* and *Stellateness 2* inside the region are included in the subset.

Region Size Some features depend on the size of the lesion, like the contrast feature. Bigger lesions have a higher contrast than smaller lesions. This morphological feature captures this difference.

Compactness Compactness represents the roughness of an object's boundary relative to its area. This feature is included because benign masses often have a round or oval shape compared to a more irregular shape of malignant masses. Compactness (C) is defined as the ratio of the squared perimeter (P) to the area (A), i.e.,

$$C = \frac{P^2}{A}$$

The smallest value of compactness is $C = \frac{(2\pi r)^2}{\pi r^2} = 4\pi$ which is for a circle. For more complex shapes, the compactness becomes larger. In our dataset this feature is normalized by dividing the compactness by 4π.

Linear Texture Normal breast tissue often has different texture characteristics than tumor tissue. Therefore Karssemeijer and te Brake[44] developed a texture feature that

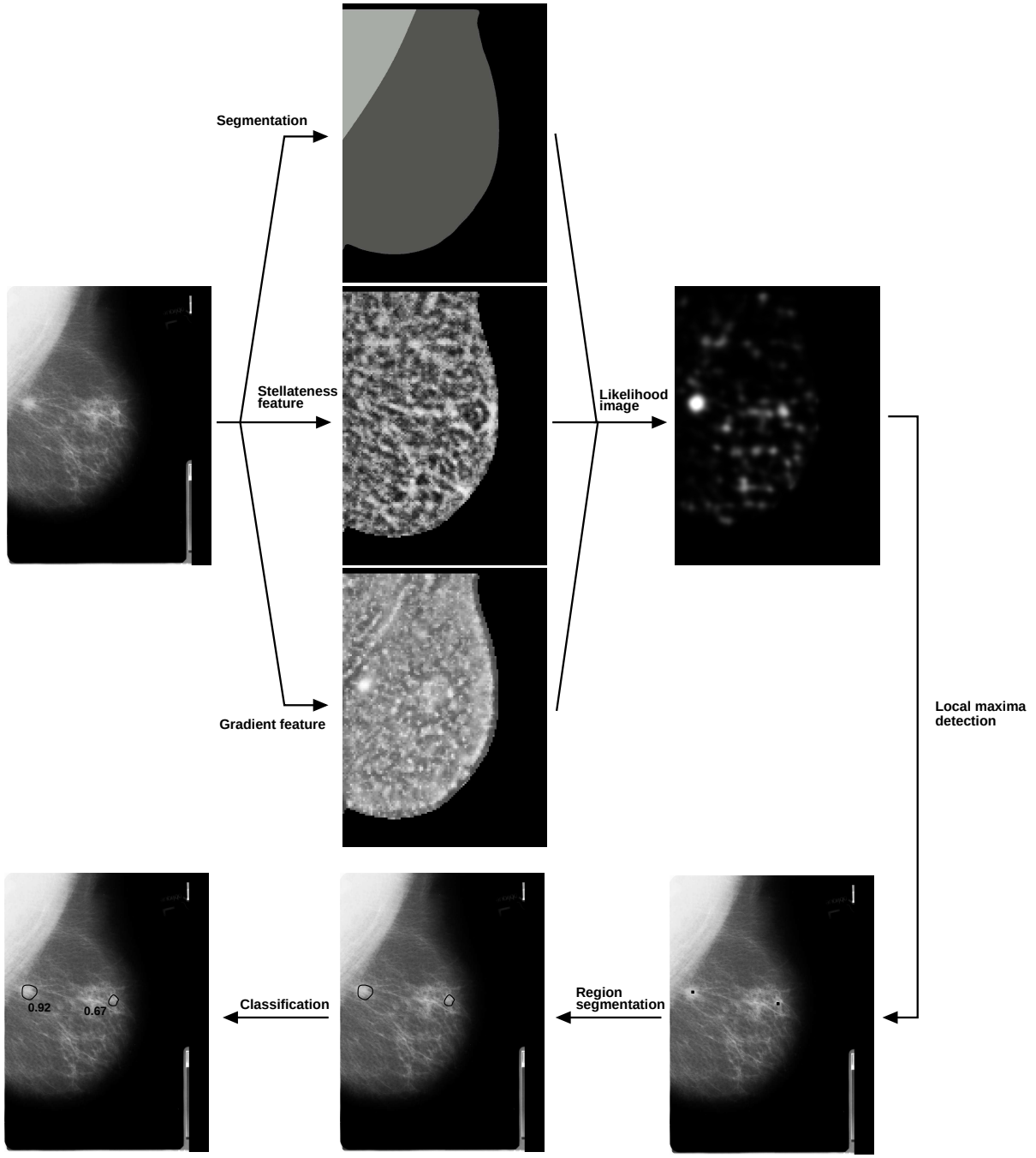


Figure 2.1: Schematic overview of the CAD scheme employed in this paper. First the mammogram is segmented into breast tissue, background tissue and the pectoral muscle. We then calculate at each location two stellateness features for the detection of spiculation and two gradient features for the detection of a focal mass. A neural network classifier combines these features into a likelihood of a mass at that location, resulting in a likelihood image. The most suspicious locations on the likelihood image (bright spots) are selected and used as seed points for the region segmentation. After that, features are calculated for each segmented region. Finally a second classifier combines these features into a malignancy score that represents the likelihood that the region is malignant.

represents presence of linear structures inside the segmented region. Malignant lesions tend to have less linear structures than normal tissue or benign lesions.

Relative Location The relative location of a lesion is important since more malignancies develop in the upper outer quadrant[150] of the breast toward the armpit. Therefore, some features have been constructed that represent the relative location of a lesion using a new coordinate system[77]. This internal coordinate system is different for MLO and CC views. In MLO views the pectoral edge is used as the y-axis. The x-axis is determined by drawing a line perpendicular to the y-axis where the distance between the y-axis and the breast boundary is maximum. We assume that the end of this line is close to the nipple. In CC views the chest wall is used as y-axis. In this internal coordinate system we calculate the x- and y-location of the centre of the segmented region and normalize with the effective radius of the breast $r = \sqrt{\frac{A}{\pi}}$, where A is the size of the segmented breast area. In this way, positions of cancers in different mammograms can be compared.

Maximum Second Order Derivative Correlation This border feature indicates the smoothness of the contour and is especially useful to discriminate between benign and malignant lesions. Most benign lesions have a well-defined contour and the margins of these lesions are sharply confined with a sharp transition between the lesion and the surrounding tissue which indicates that there is no infiltration[77].

Contrast Regions with high contrast or a higher intensity than other similar structures in the image are more likely to be a mass since tumor tissue on average absorbs more X-rays than fat and also slightly more than glandular tissue. The distance measure we used to indicate differences in contrast is the squared difference in intensity between the segmented region and its surround, divided by both standard deviations,

$$\frac{(\overline{Y}(R) - \overline{Y}(S))^2}{\sigma_Y(R) + \sigma_Y(S)}$$

where R is the set of pixels in the segmented region, S is the set of pixels in the surroundings of the segmented region. $\overline{Y}(X)$ is the mean grey level of the pixels in set X, and $\sigma_Y(X)$ is the grey level standard deviation of the pixels in set X.

Number of Calcifications The presence of clustered microcalcifications is one of the most important signs of cancer on a mammogram. They occur in about 90% of the non-invasive cancers. Therefore we include a feature representing the number of calcifications found in the segmented region.

	Mean	Std dev	Min	Max	Skewness	Kurtosis
Benign (cases: 258)						
Stellateness 1	1.1256	0.1710	0.7800	2.1400	2.3002	13.4307
Stellateness 2	1.0241	0.1160	0.8300	2.1900	4.7815	44.8670
Stellateness 1 Mean	1.1189	0.1316	0.8600	1.5630	0.8565	3.6986
Stellateness 2 Mean	1.0215	0.0713	0.8380	1.2990	0.5482	3.6256
Region Size	0.4070	0.3915	0.0200	3.4510	3.0272	17.9799
Contrast	0.5502	0.2558	0.1260	2.0110	1.9986	9.8575
Compactness	1.2141	0.0906	1.0470	1.5600	0.9308	3.8448
Linear Texture	0.1750	0.1444	0.0130	1.0240	2.2365	10.1391
Relative Location X	0.6705	0.3024	-0.0670	1.5470	0.0470	2.7819
Relative Location Y	0.2160	0.4262	-0.9680	1.2990	-0.2289	2.4769
Max. 2nd order Drv Corr.	0.6800	0.1008	0.4520	0.9060	0.0436	2.3011
Number of Calcifications	0.7871	2.6723	0.0000	19.0000	3.8831	19.2635
Malignant (cases: 274)						
Stellateness 1	1.2273	0.1730	0.8200	1.7300	0.5060	3.0005
Stellateness 2	1.0827	0.0965	0.7900	1.3500	0.1468	2.8634
Stellateness 1 Mean	1.2357	0.1736	0.8290	1.7740	0.6844	3.1281
Stellateness 2 Mean	1.0868	0.0946	0.8530	1.4140	0.4533	3.0175
Region Size	0.4471	0.3272	0.0160	1.8040	1.2728	4.4259
Contrast	0.6272	0.2777	0.0110	1.5090	0.7688	3.2074
Compactness	1.2111	0.0983	1.0410	1.7080	1.5022	6.3482
Linear Texture	0.1578	0.1161	0.0040	0.9490	2.2258	11.5829
Relative Location X	0.6130	0.3046	-0.0710	1.3080	0.0140	2.3298
Relative Location Y	0.2080	0.4449	-0.9770	1.2180	-0.2483	2.7594
Max. 2nd order Drv Corr.	0.6354	0.0951	0.4040	0.9320	0.1608	2.9336
Number of Calcifications	2.0645	6.7471	0.0000	50.0000	4.4524	25.7707

Table 2.1: Statistics of benign and malignant cases in the used dataset

2.2.3 Statistical analysis

For every feature the first four moments of the distribution of feature values in the dataset have been computed. These are shown in Table 2.1. The third moment, skewness, is a measure of the lack of symmetry. The skewness for a normal distribution is zero, and any near-symmetric data should have a skewness near zero. The fourth moment, also called kurtosis, is a measure of whether the data are peaked or flat relative to a normal distribution. The kurtosis for a standard normal distribution is three.

Combining the 12 features of the MLO views with the 12 features of the corresponding CC views gives a total of 24 features per case. The continuous output of the classifier is analyzed using ROC methodology, using the LABROC program[151] of Metz et al. The statistical significance of the difference between ROC curves was tested using the CLABROC program[152] of Metz et al. The classifiers were trained and tested using a k-fold cross-validation test method (k=10), in which each of 10 different combinations of training and test data sets included 244 and 27 cases, respectively. For each test partition, the classification accuracy was evaluated as the area A_z under the ROC curve.

2.2.4 Naive Bayesian classifier

The naive Bayesian classifier (Figure 2.2) is a Bayesian network with a limited topology[153] applicable to learning tasks where each instance is described by a conjunction of feature values and a class value. To learn the Bayesian network a set of training examples has to be provided. Classification using this Bayes' probability model is done by picking the most probable hypothesis which is also known as the *maximum a posteriori*. The corresponding classifier function can be defined as follows:

$$C_{MAP} = \arg\max_{c_j \in C} \; P(c_j | f_1, f_2, \ldots, f_n) \tag{2.1}$$

where $\{f_1, f_2, \ldots, f_n\}$ is the set of feature values that decribe the new instance, and C_{MAP} is the most probable hypothesis. Using Bayes theorem, Equation 2.1 can be rewritten as follows:

$$\begin{aligned} C_{MAP} &= \arg\max_{c_j \in C} \; \frac{P(c_j)P(f_1, f_2, \ldots, f_n | c_j)}{P(f_1, f_2, \ldots, f_n)} \\ &= \arg\max_{c_j \in C} \; P(c_j)P(f_1, f_2, \ldots, f_n | c_j) \end{aligned} \tag{2.2}$$

Using training data the two terms $P(c_j)$ and $P(f_1, f_2, \ldots, f_n | c_j)$ have to be calculated. The *class prior probability* $P(c_j)$ can be easily estimated by counting the frequency of occurence of the class value c_j in the training data. However, estimating the different $P(f_1, f_2, \ldots, f_n | c_j)$ terms is difficult and is only possible if a huge set of training data is available. To dramatically simplify the classification task we can use the following simplifying assumption: each feature f_i is conditionally independent of every other feature f_j for $i \neq j$. This fairly strong assumption of independence leads to the name naive Bayes, with the assumption often being naive in that, by making this assumption, the algorithm does not take into account dependencies that may exist. By using the conditionally independence assumptions we can express Equation 2.2 as:

$$C_{MAP} = \arg\max_{c_j \in C} \; P(c_j) \prod_{i=1}^{n} P(f_i | c_j) \tag{2.3}$$

The model in this form is much more manageable, since it factors into a so-called *class prior probability* $P(c_j)$ and independent probability distributions $P(f_i | c_j)$. These class conditional probabilities $P(f_i | c_j)$ can be calculated separately for each variable which reduces complexity enormously. Even with such strong simplifying assumptions, it does not seem to greatly affect the posterior probabilities, especially in regions near the decision boundaries which leaves the classification task unaffected. Some papers show that such naive Bayesian classifiers yield surprisingly powerful classifiers[154].

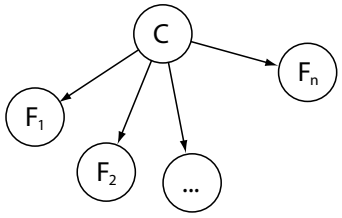


Figure 2.2: A graphical representation of a naive Bayesian classifier

2.2.5 Support vector machines

The SVM algorithm has been introduced by Cortes and Vapnik[140] for solving classification tasks and have been successfully applied in various areas of research. The basic idea of SVM is that it projects datapoints from a given two-class training set in a higher dimensional space and finds an optimal hyperplane. The optimal one is the one that separates the data with the maximal margin. SVMs identify the datapoints near the optimal separating hyperplane which are called support vectors. The distance between the separating hyperplane and the nearest of the positive and negative datapoints is called the margin of the SVM classifier. The separating hyperplane is defined as

$$D(x) = (w \cdot x) + b \tag{2.4}$$

where x is a vector of the dataset mapped to a high dimensional space, and w and b are parameters of the hyperplane that the SVM will estimate. The nearest datapoints to the maximum margin hyperplane lie on the planes

$$\begin{aligned} (w \cdot x) + b &= +1 \quad \textit{for } y = +1 \\ (w \cdot x) + b &= -1 \quad \textit{for } y = -1 \end{aligned} \tag{2.5}$$

where $y = +1$ for class ω_1 and $y = -1$ for class ω_2. The width of the margin is given by $m = \frac{2}{||w||}$. Computing w and x is then the problem of finding the minimum of a function with the following constraints:

$$\begin{aligned} \textit{minimize} \qquad & m(w) = \frac{1}{2}(w \cdot w) \\ \textit{subject to constraints} \qquad & y_i[w \cdot x_i + b] \geq 1 \end{aligned} \tag{2.6}$$

In its simplest form, a SVM attempts to find a linear separator, as shown in Figure 2.3. In practice however, there may be no good linear separator of the data. In that case, SVMs can project the dataset to a significant higher dimensional feature space to make the separation easier, using a kernel function to produce separators that are non-linear. Unfortunately there is no theory about deciding which kernel is the best[155].

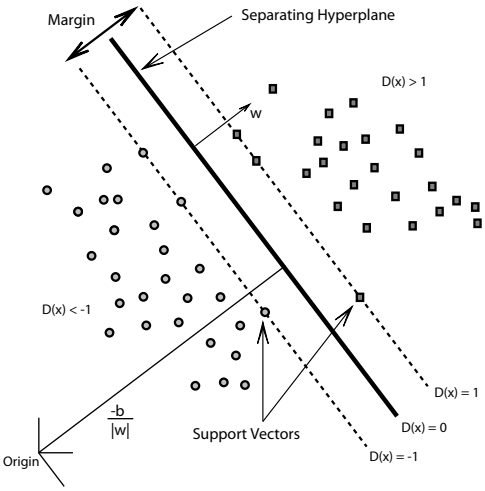


Figure 2.3: Linear separating hyperplanes for the separable case.

2.2.6 Preprocessing: Manly transformation

Many Bayesian learning algorithms that deal with continuous nodes, including the learning algorithms in Kevin Murphy's Bayesian Networks Toolbox[156], are based on the assumption that the features are normally distributed. Unfortunately, most of the image features we use do not follow a normal distribution. We used Manly's exponential transformation to make the non-normal data resemble normal data by reducing skewness, which is a transformation from y to $y^{(\lambda)}$ with parameter λ. This transform is most effective if the probability distribution of a feature can be described as a function which contains powers, logarithms, or exponentials. The transform is given by:

$$y^{(\lambda)} = \begin{cases} \frac{e^{\lambda y}-1}{\lambda} & \text{if } \lambda \neq 0 \\ y & \text{if } \lambda = 0 \end{cases} \tag{2.7}$$

The assumption made by this transformation is that $y^{(\lambda)}$ follows a normal linear model with parameters β and σ^2 for some value of λ. Given a value of λ, we can estimate the linear model parameters β and σ^2 as usual, except that we work with the transformed variable $y^{(\lambda)}$ instead of y. To select an appropriate transformation we need to find the optimal value of λ using an optimization criteria. We used a technique based on the normal probability plot. The data is plotted against a theoretical normal distribution in such a way that the points should form an approximate straight line if the data is normal distributed. Deviations of this straight line mean that the data is less normally distributed. We can use that property to plot the correlation coefficient of the normality plot against a range of λ's. The lambda resulting in the largest correlation

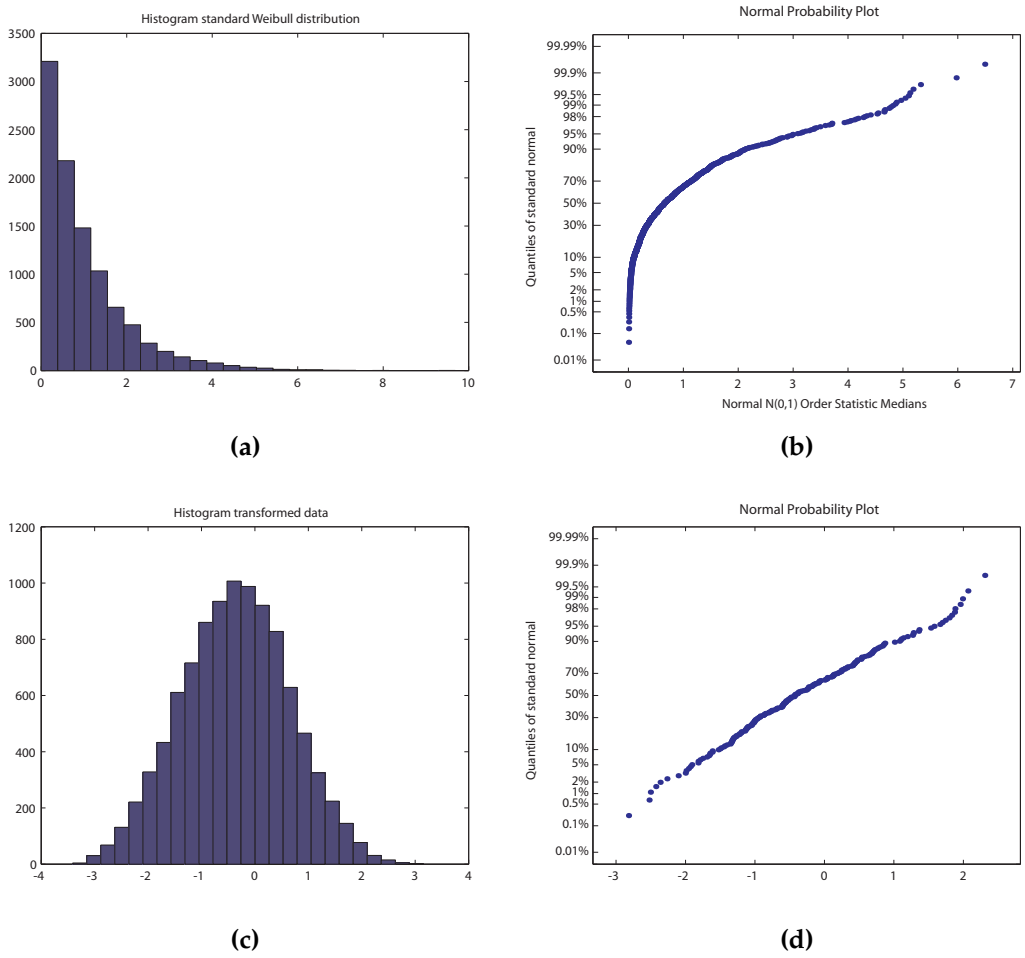


Figure 2.4: An example Manly transformation: (a) histogram of a feature that is Weibull distributed, (b) normality plot of the feature, (c) histogram of the transformed feature, and (d) normality plot of the transformed feature

coefficient is chosen.

2.2.7 Preprocessing: principal component analysis

One might think that the use of more features will automatically improve the classification power of the classifier. However the number of samples needed to train a classifier with a certain level of accuracy increases exponentially with the number of features. Therefore, we used principal component analysis[157] as a preprocessing technique to reduce the dimensionality of our dataset. The assumption made in PCA is that most of the information is carried in the variance of the features: the higher the variance in one dimension (feature), the more information is carried by that feature. The general idea is therefore to preserve the most variance in the data using the least number of dimensions. One of the major drawbacks of PCA is that it is an unsupervised

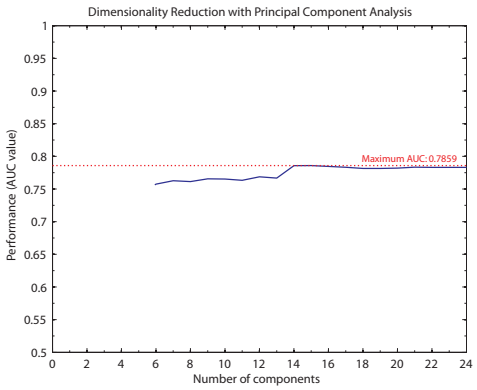


Figure 2.5: Case based performance naive Bayes classifier after dimensionality reduction with PCA, averaged over 5 runs.

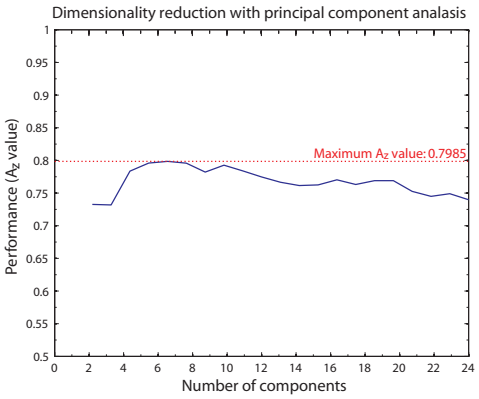


Figure 2.6: Case based performance SVM classifier with radial kernel function after dimensionality reduction with PCA, averaged over 5 runs.

algorithm, i.e., it does not take the class label in account. It can therefore eliminate a dimension that is best for discriminating positive from negative cases.

2.3 Results

The dataset we used contained a lot of features that were highly skewed and therefore did not follow a normal distribution. The learning algorithms in Murphy's BNT toolbox[156] for Bayesian networks with continuous nodes, assume that within each state of the class the observed continuous features follow a normal distribution. These continuous nodes have therefore two parameters per class, mean and variance, to represent the characteristics of the training data. We evaluated the classification performance of the naive Bayes classifier after applying the Manly transformation on the dataset. The *Stellateness Mean* and the *Maximum Second Order Derivative Correlation* features are approximately normal distributed in their original form and did not perform well when transformed. We chose therefore to not transform these features. Also the *Number of Calcifications* feature was not a useful candidate to transform, because of its discrete nature. Statistical information about the transformed dataset can be found in Table 2.2.

The calculated area under the ROC curve (A_z value) of the Bayesian classifier without transforming the dataset was 0.767. After applying the Manly transformation it increased to 0.795, which is statistically significant (p=0.0002). For the SVM classifier, the Manly transformation had no noticeable effect on the performance. Comparing the performance between BNs and SVMs using the transformed dataset showed that the difference was not statistically significant (p=0.78).

Additionally, we evaluated the classification performance of the naive Bayesian

	Mean	Std dev	Min	Max	Skewness	Kurtosis
All cases (cases: 542)						
Stellateness 1	0.5102	0.0229	0.4351	0.5799	0.0000	3.1743
Stellateness 2	0.4077	0.0095	0.3739	0.4495	0.0000	3.7745
Stellateness 1 Mean	1.1790	0.1653	0.8290	1.7740	0.8638	3.5949
Stellateness 2 Mean	1.0551	0.0902	0.8380	1.4140	0.6548	3.4349
Region Size	0.2148	0.0878	0.0157	0.3706	0.0000	1.9075
Contrast	0.3599	0.0924	0.0109	0.5939	0.0000	2.7383
Compactness	0.2070	0.0002	0.2063	0.2076	0.0000	2.4822
Linear Texture	0.0931	0.0381	0.0040	0.1725	0.0000	2.3284
Relative Location X	0.6312	0.2974	-0.0711	1.5016	0.0000	2.5448
Relative Location Y	0.2404	0.4554	-0.8736	1.5173	0.0000	2.5943
Max. 2nd order Drv Corr.	0.6571	0.1004	0.4040	0.9320	0.1290	2.5924
Number of Calcifications	1.4446	5.2255	0.0000	50.0000	5.4429	39.8079
Benign (cases: 263)						
Stellateness 1	0.5029	0.0219	0.4351	0.5799	0.2717	4.1962
Stellateness 2	0.4047	0.0091	0.3806	0.4495	0.5072	5.8195
Stellateness 1 Mean	1.1189	0.1316	0.8600	1.5630	0.8565	3.6986
Stellateness 2 Mean	1.0215	0.0713	0.8380	1.2990	0.5482	3.6256
Region Size	0.2048	0.0882	0.0195	0.3706	0.1791	1.9413
Contrast	0.3463	0.0865	0.1140	0.5939	0.2547	2.8499
Compactness	0.2070	0.0002	0.2063	0.2075	-0.1018	2.4716
Linear Texture	0.0946	0.0396	0.0125	0.1725	-0.0221	2.2135
Relative Location X	0.6601	0.2946	-0.0671	1.5016	0.0159	2.7614
Relative Location Y	0.2437	0.4457	-0.8665	1.5173	-0.0159	2.4540
Max. 2nd order Drv Corr.	0.6800	0.1008	0.4520	0.9060	0.0436	2.3011
Number of Calcifications	0.7871	2.6723	0.0000	19.0000	3.8831	19.2635
Malignant (cases: 279)						
Stellateness 1	0.5171	0.0217	0.4456	0.5636	-0.2453	2.9714
Stellateness 2	0.4106	0.0090	0.3739	0.4301	-0.4677	3.4270
Stellateness 1 Mean	1.2357	0.1736	0.8290	1.7740	0.6844	3.1281
Stellateness 2 Mean	1.0868	0.0946	0.8530	1.4140	0.4533	3.0175
Region Size	0.2242	0.0864	0.0157	0.3678	-0.1658	1.9722
Contrast	0.3728	0.0960	0.0109	0.5640	-0.2511	2.8418
Compactness	0.2070	0.0002	0.2063	0.2076	0.0943	2.5107
Linear Texture	0.0917	0.0365	0.0040	0.1722	0.0067	2.4444
Relative Location X	0.6040	0.2975	-0.0711	1.2754	-0.0094	2.3235
Relative Location Y	0.2372	0.4643	-0.8736	1.4087	0.0149	2.6997
Max. 2nd order Drv Corr.	0.6354	0.0951	0.4040	0.9320	0.1608	2.9336
Number of Calcifications	2.0645	6.7471	0.0000	50.0000	4.4524	25.7707

Table 2.2: Statistics of benign and malignant cases after transformation.

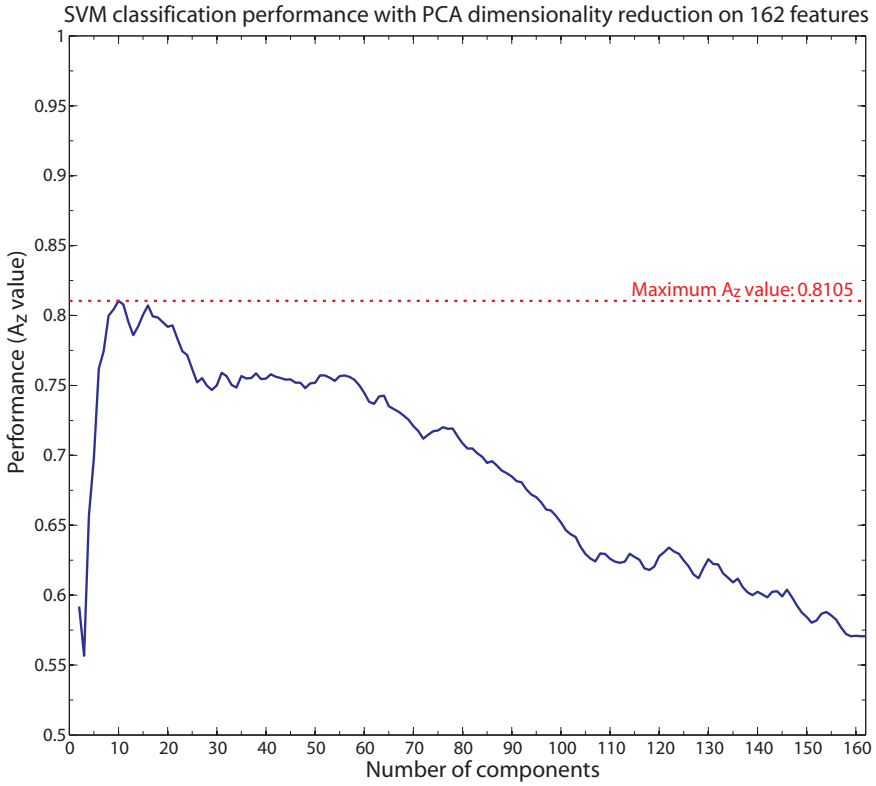


Figure 2.7: Case based performance SVM classifier with radial kernel function after dimensionality reduction of all features (81 per view) with PCA, averaged over 5 runs.

and SVM classifier after applying dimensionality reduction on our dataset. Figure 2.5 shows the classification performance of the naive Bayesian classifier, where horizontally the number of principal components is plotted and vertically the area under the ROC curve. The principal component vectors were calculated using the training set only. These principal component vectors are then used to transform both the training and test set. The best result was obtained with 14 principal components. The performance remained almost constant when adding more dimensions. With SVMs the best result was obtained with only 6 principal components and decreased gradually if more components were added which is shown in Figure 2.6. The difference in classification performance between BNs and SVMs was statistically insignificant (p=0.11) when we used the optimal number of principal components for the classifier. In an additional experiment we trained a SVM on all the available features (81 per view). This led to the classification results shown in Figure 2.7. The maximum performance was reached in 10 components ($A_z = 0.811$) but this was not significantly higher than the maximum performance obtained in the experiment with the subset of the 12 most important features ($A_z = 0.793$).

2.4 Conclusions

We performed a study to compare two state-of-the-art classification techniques characterizing masses as either benign or malignant. We evaluated the effectiveness of dimension reduction and normal distribution transformation in improving the classification accuracy. The Manly transformation method significantly improved classification accuracy of the naive Bayesian classifier. We believe that this is due the fact that, by transforming the distribution of the non-normal data to a distribution closer to normal, the assumptions of the naive Bayesian classifier are violated less. We also found that this transformation does not work for all data, i.e., transforming features that were already approximately normal within their class. We believe that by selecting one gamma for Manly's transformation, without looking to the class label, can negatively effect the binormal distribution (i.e., two normal distributions: one for benign and another for malignant cases) of the *Stellateness Mean* features. For the SVM classifier, the data does not need to be normally distributed which explains why this transformation did not have effect on the performance of the SVM classifier. After transformation, the difference in performance of the SVM classifier and the naive Bayesian classifier was not statistically significant. Bayesian networks allow incorporating background knowledge, which may be exploited to improve their performance in the future. Despite the major drawback of principal component analysis, i.e., it can eliminate a dimension that is good for discriminating positive cases from negative cases, this unsupervised dimension reduction algorithm improved the classification accuracy of both classifiers. The performance of the two classifiers after applying PCA was very similar, with no statistical differences in the area under the ROC curve.

Improved mammographic CAD performance using multi-view information

3

Marina Velikova, Maurice Samulski, Peter Lucas, and Nico Karssemeijer

Original title: Improved mammographic CAD performance using multi-view information: A Bayesian network framework

Published in: Physics in Medicine and Biology 2009;54(5):1131-1147

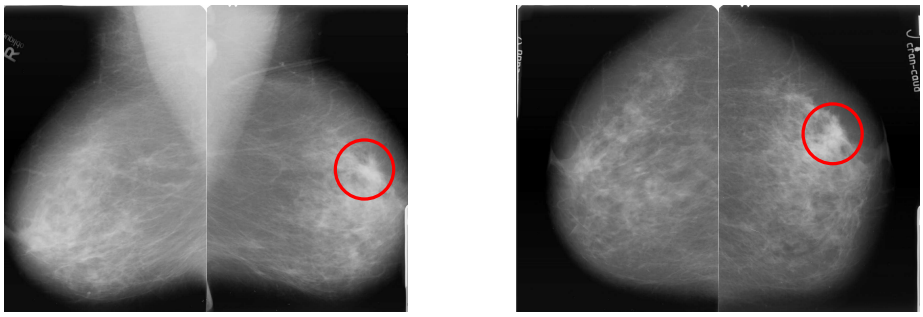
Abstract

Mammographic reading by radiologists requires the comparison of at least two breast projections (views) for the detection and the diagnosis of breast abnormalities. Despite their reported potential to support radiologists, most mammographic computer-aided detection (CAD) systems have a major limitation: as opposed to the radiologist's practice, computerized systems analyze each view independently. To tackle this problem, in this paper, we propose a Bayesian network framework for multi-view mammographic analysis, with main focus on breast cancer detection at a patient level. We use causal independence models and context modelling over the whole breast represented as links between the regions detected by a single-view CAD system in the two breast projections. The proposed approach is implemented and tested with screening mammograms for 1063 cases of whom 385 had breast cancer. The single-view CAD system is used as a benchmark method for comparison. The results show that our multi-view modelling leads to significantly better performance in discriminating between normal and cancerous patients. We also demonstrate the potential of our multi-view system for selecting the most suspicious cases.

3.1 Introduction

Breast cancer is the most common form of cancer among women world-wide and its early detection does improve the chances of successful treatment and recovery[158]. Therefore, many countries have introduced breast cancer screening programs with periodic mammographic examinations in asymptomatic women. In contrast to the clinical situation, in the screening setting the detected lesions are usually small and due to the breast compression they are sometimes difficult to observe in both views. In other words, while the correct detection and location of a cancerous region is important, in breast cancer screening the crucial decision based on the mammographic exam is whether or not it is likely that a woman has breast cancer, and if the answer is positive, is referred to the clinic for further examination.

A screening mammographic examination usually consists of four images, corresponding to each breast scanned in two views–mediolateral oblique (MLO) view and craniocaudal (CC) view (see Figure 3.1). The MLO projection is taken under 45° angle and shows part of the pectoral muscle. The CC projection is a top-down view of the breast. In reading mammograms, radiologists judge whether or not a lesion is present by comparing both views and breasts. The general rule is that a lesion is to be observed in both views.



a) b)

Figure 3.1: a) MLO and b) CC views of a right and left breast of a patient. The circle depicts a (cancerous) lesion in the left breast

To guarantee high detection rates, independent double reading by two radiologists is a widely used standard in breast cancer screening. Due to its complexity and the variability in human performance, however, mammographic reading and decision making appear to be difficult tasks. Radiologists are usually confronted with two main problems in the mammographic analysis: (i) perceptual oversight where an abnormality is present, but is missed and (ii) interpretation failure where an abnormality is seen

but its significance is misinterpreted. There are two main types of abnormalities: microcalcifications and masses. In this work, we deal with the second, more frequently occurring type. There is strong evidence that for masses misinterpretation is a more common cause of missing cancers in screening than perceptual oversight.

In an attempt to support radiologists in overcoming these problems, a large number of mammographic computer-aided detection (CAD) systems have been developed and tested in the past twenty years. Essentially, the working principle of current CAD systems comprises a multi-stage process based on identification of regions of interest using image processing and pattern recognition techniques, extraction of a feature vector for each of these regions and classification of the regions as cancerous (abnormal) based on supervised learning techniques such as neural networks.

Despite the reported evidence about their potential benefit, most CAD methods suffer from certain limitations due to the uncertainty inherent in the domain. For example, misclassification can arise between a region of interest and its extracted feature vector or lack of separability between regions of interest that have similar features. One reason for these problems is that, opposite to the radiologist's practice, most computerized systems are based on a single-view principle where each view and the regions within a view are analyzed independently. Hence, the multi-view and multi-region dependencies in the breast are ignored and the breast cancer detection can be obscured. As a result, such systems perform worse than the human experts, which limits their practical application and usability.

To tackle these problems statistical modelling of the domain can be applied to the automatic detection process. In this paper, we propose a Bayesian network framework for exploiting multi-view dependencies for the analysis of screening mammograms. Given the goal of screening programs, we focus primarily on the breast cancer detection at a patient level, rather than on the location of the cancer in the mammogram. The main idea of our methodology lies in combining the information available as detected regions from a single-view CAD system in MLO and CC to obtain a single likelihood measure for a patient being cancerous. In comparison to previous methods, we can outline a number of advantages of our probabilistic framework:

- Handling noise and missing information: specifying and learning the network parameters in a probabilistic manner allows uncertain information to be incorporated based on the values of all the non-missing variables.

- Incorporating domain knowledge: unlike black-box approaches such as neural networks, our framework captures explicitly view dependencies through the Bayesian network structure and the definition of the conditional probability tables.

- Using context information over the whole breast: breast classification is done on the basis of simultaneous consideration of the regions automatically detected in each breast view and their links to the other view of the same breast.

We adopt the following terminology from the breast cancer domain throughout this paper. By *lesion* we refer to a physical cancerous object detected in a patient (the circle in Figure 3.1). We call a contoured area on a mammogram a *region* (for example, marked manually by a human or detected automatically by a CAD system). A region can be true positive (TP), i.e., correct detection of the lesion (cancer) or false positive (FP). A region detected by a CAD system is described by a number of continuous (real-valued) *features* (e.g., size, location, contrast). By *link* we denote matching (established correspondence) between two regions in MLO and CC views, respectively. The term *case* refers to a patient who has undergone a mammographic exam. The most recent case for a patient is called *current* whereas the previous case(s) are *prior(s)*.

The remainder of the paper is organized as follows. In the next section we briefly review previous research in multi-view breast cancer detection. In Section 3.3 we describe the general problem of multi-view detection, introduce basic definitions related to Bayesian networks and then we present a general Bayesian network framework for multi-view detection. The proposed approach is evaluated on an application of breast cancer detection using actual screening data. The evaluation procedure and the results are presented in Section 3.4. Conclusions and directions for extension of our model are given in Section 3.5.

3.2 Previous research

A number of previous works deal with the problem of automatic multi-view breast cancer detection on mammograms. Good et al.[159] proposed a probabilistic method for true matching of lesions detected in both views, based on Bayesian network and multi-view features. The results from experiments demonstrate that their method can significantly distinguish between true and false positive links of regions. Van Engeland et al.[160] describe another linking method based on Linear Discriminant Analysis (LDA) classifier and a set of view-link features to compute a correspondence score for every possible region combination. For every region in the original view the region in the other view with the highest correspondence score is selected as the corresponding candidate region. The proposed approach demonstrates an ability to discriminate between true and false links. Van Engeland and Karssemeijer[51] extended this matching approach by building a cascaded multiple-classifier system for reclassifying the region level of suspiciousness of an initially detected region based on the linked candi-

date region in the other view. Experiments have shown that the lesion-based detection performance of the two-view detection system is significantly better than that of the single-view detection method.

Paquerault et al.[53] also consider established correspondence between suspected regions in both views to improve lesion detection. LDA is used to classify each object pair as true or false. By combining the resulting correspondence score with its one-view detection score the lesion detection improves and the number of false positives reduces. In this study, the authors also report improvement in the case-based performance (fraction of TP cases where a case is TP, if cancer is found in MLO *or* CC view) based on multi-view information, especially for cases where the lesion has been detected in both views.

In two recent studies, Sun et al.[161] and Qian et al.[162] also demonstrate the superior performance of a multi-view CAD system over its single-view counterpart. The approach consists of multiple steps starting with advanced single-view image processing for region segmentation, followed by multi-view feature extraction and final classification of the detected regions of interest based on neural networks with Kalman filtering. Using iterative processing between the single- and multi-view stages, the authors show a reduction at the false positive rates of masses per image as well as an increase at the case-based detection rate.

However, in all these works the main focus is on improving the localized detection of breast cancer, mostly for prompting purposes, rather than the detection at a case level. Therefore, the likelihood for cancer in a case is often determined by the region with the maximum likelihood. In contrast, in the current study we aim at building a CAD system that discriminates well between normal and cancerous cases–the ultimate goal of breast cancer screening programs–by considering all available information (in terms of regions) in a case. In the next section, we describe such a system based on a probabilistic methodology and we demonstrate its practical potential on a case study.

3.3 Bayesian multi-view detection

3.3.1 Problem Description

In multi-view medical imaging, two-dimensional (2D) projections of the organ(s) of interest (e.g. breast) are acquired from two or more viewing angles. The objective of the multi-view detection then is to determine whether or not the object has certain characteristics (e.g., being cancerous) by establishing correspondences between the 2D image characteristics of regions (subparts) in multiple object views (projections). Figure 3.2 depicts the general multi-view detection scheme.

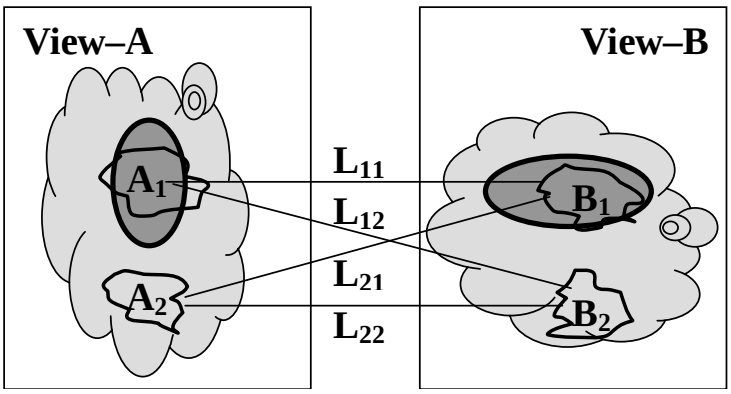


Figure 3.2: Schematic representation of multi-view analysis of a physical object with automatically detected regions

We have a physical object referring to an organ (displayed as a gray cloud), which is projected in two views, *View-A* and *View-B*. Suppose we have a cancerous physical subpart of the object represented by the ovals in both projections; hence, the whole object is cancerous. In both views an automatic single-view system detects potential cancerous regions described by a number of real-valued extracted features. In the figure regions A_1 and B_1 are correct detection of the cancerous physical subpart, i.e., these are TP regions whereas A_2 and B_2 are FP regions. Since we deal with projections of the same physical object we introduce *links* (L_{ij}) between the detected regions in both views, A_i and B_j. Every link has a class (label) $L_{ij} = \ell_{ij}$ defined as follows

$$\ell_{ij} = \begin{cases} true & \text{if } A_i \text{ or } B_j \text{ are TP,} \\ false & \text{otherwise.} \end{cases} \tag{3.1}$$

This definition allows us to maintain information about the presence of cancer even if there is no cancer detection in one of the views. A binary class C with values of $true$ (presence of cancer) and $false$ for region, view or the whole object (organ) is assumed to be provided by pathology or a human expert.

In any case, multiple views corresponding to the same cancerous part contain correlated characteristics whereas views corresponding to normal parts tend to be less correlated. For example, in mammography an artifactual density might appear in one view due to the superposition of normal tissue whereas it disappears in the other view. To account for the interaction between the breast projections, in this paper we develop a Bayesian network framework for mammographic analysis. The power of Bayesian networks lies in their ability to (i) explicitly and efficiently encode causal dependences in a domain and (ii) model and reason about uncertainty in a probabilistic fashion. This makes them a suitable modelling tool for the multi-view detection problem. The

next section gives some general background about Bayesian networks.

3.3.2 Bayesian Networks

Consider a finite set $\mathbf{U}$ of random variables, where each variable U in $\mathbf{U}$ takes on values from a finite domain $dom(U)$. Let P be a joint probability distribution of $\mathbf{U}$ and let $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ be disjoint subsets of $\mathbf{U}$. We say that $\mathbf{X}$ and $\mathbf{Y}$ are conditionally independent given $\mathbf{Z}$, denoted by $\mathbf{X} \perp\!\!\!\perp_P \mathbf{Y} \mid \mathbf{Z}$, if for all $\mathbf{x} \in dom(\mathbf{X})$, $\mathbf{y} \in dom(\mathbf{Y})$, $\mathbf{z} \in dom(\mathbf{Z})$, the following holds:

$$P(\mathbf{x} \mid \mathbf{y}, \mathbf{z}) = P(\mathbf{x} \mid \mathbf{z}), \text{ whenever } P(\mathbf{y}, \mathbf{z}) > 0.$$

In short, we have $P(\mathbf{X} \mid \mathbf{Y}, \mathbf{Z}) = P(\mathbf{X} \mid \mathbf{Z})$.

A Bayesian network is defined as a pair $\mathrm{BN} = (G, P)$ where G is an acyclic directed graph (ADG) $G = (V, E)$ with a set of nodes V corresponding to the random variables in $\mathbf{U}$ and a set of edges (arcs) $E \subseteq (V \times V)$ corresponding to direct causal relationships between the variables. We say that G is an *I–map* of P if any independence represented in G, denoted by $A \perp\!\!\!\perp_G B \mid C$ with $A, B, C \subseteq V$ mutually disjoint sets of nodes, is satisfied by P, i.e.,

$$A \perp\!\!\!\perp_G B \mid C \quad \Longrightarrow \quad \mathbf{X}_A \perp\!\!\!\perp_P \mathbf{X}_B \mid \mathbf{X}_C,$$

where A, B and C are sets of nodes of the ADG G and $\mathbf{X}_A$, $\mathbf{X}_B$ and $\mathbf{X}_C$ are the corresponding sets of random variables. The acyclic directed graphical part of a Bayesian network G is by definition an I–map of the associated joint probability distribution P. A Bayesian network BN offers a compact representation of the joint probability distribution P in terms of local *conditional probability distributions* (*CPDs*), or, in the discrete case, in terms of *conditional probability tables* (*CPTs*), associated to the individual nodes. The conditional probability distributions are usually more compact than in the general case, as they take into account the conditional independence information represented by the ADG. For a more detailed recent description of Bayesian networks, the reader is referred to[163].

Causal Independence Models

It is known that the number of probabilities in a CPT for a certain variable grows exponentially in the number of parents in the ADG. Therefore it is often infeasible to define the complete CPT for variables with many parents. One way to specify interactions among statistical variables in a compact fashion is offered by the notion of *causal independence*[164]. Causal independence arises in cases where multiple causes (parent nodes) lead to a common effect (child node). Here we present the formal definition

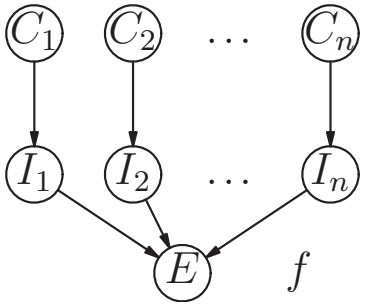


Figure 3.3: Causal-independence model

of the notion of causal independence as given in the article from Lucas 2005[165]. The general structure of a causal-independence model is shown in Figure 3.3; it expresses the idea that causes $C_1, \ldots, C_n$ influence a given common effect E through intermediate variables $I_1, \ldots, I_n$; the intermediate variable I_k is considered to be a contribution of the cause variable C_k to the common effect E. The *interaction function* f represents in which way the intermediate effects I_k, and indirectly also the causes C_k, interact. This function f is defined in such way that when a relationship between the I_k's and $E = true$ is satisfied, then it holds that $f(I_1, \ldots, I_n) =$ *true*; otherwise, it holds that $f(I_1, \ldots, I_n) =$ *false*. Note that each variable I_k is only dependent on its associated cause C_k and the effect variable E. Furthermore, the graph structure expresses that the effect variable E is conditionally independent of each cause C_k given the associated intermediate variable I_k.

An important subclass of causal-independence models is obtained if the deterministic function f is defined in terms of separate binary functions g_k; it is then called a *decomposable* causal-independence model[164]. Usually, all functions $g_k(I_k, I_{k+1})$ are identical for each k. Typical examples of decomposable causal-independence models are the noisy-OR[166] models, where the function g represents a logical OR. These models express that the presence of any of the causes C_k with absolute certainty will cause the effect $E = true$. A simple example of a noisy-OR model is given in the Appendix.

In our modelling framework, presented in the next section, we apply such a decomposable causal-independence model with the logical OR. Our choice is motivated by two major features of the representation of the noisy-OR models. First, from the definition of the noisy-OR model it follows that the higher the number of causes influencing the effect the higher the probability that the effect occurs. This rule is definitely applicable in the domain of breast cancer detection where the more evidence (e.g., in terms of detected regions) is added the higher the probability for cancer. Another important feature from a computational point of view is that the representation of the noisy-OR models has a linear complexity with respect to the number of causes.

3.3.3 Model Description

Our modelling scheme is based on two Bayesian networks with a hand-constructed (fixed) structure to explicitly represent the multi-view dependencies in the detection problem. Consider again the detection scheme presented in Figure 3.2. The regions A_i and B_j are generally conditionally independent given the case. However, they become dependent once we have evidence that they are the projections of the same lesion in two views. In the context of Bayesian networks, this region dependence can be modelled by (i) three nodes: two for the regions and one for the link and (ii) the so-called *v-structure* where directed arcs are drawn from the region nodes to the link node: $A_i \longrightarrow L_{ij} \longleftarrow B_j$. Such a representation of the dependence between a relation (link) and its parts (regions) has also been advocated by other researchers in the field of vision perception[167]. Note that swapping the arc direction from the link to the regions would imply that the regions are conditionally independent given the existence of a link, which contradicts our intuition.

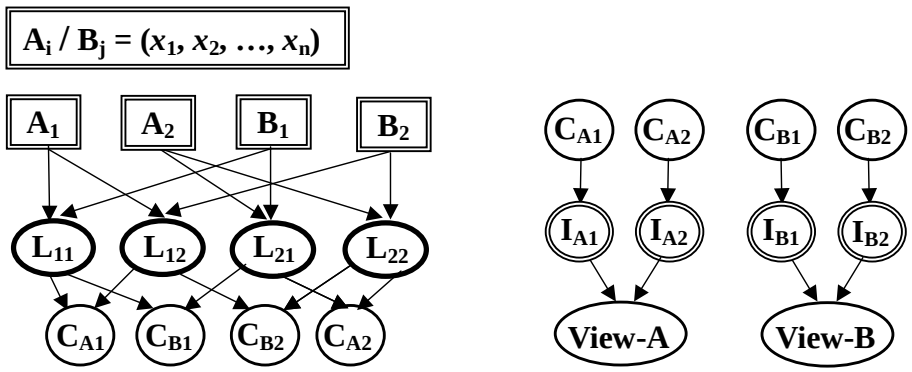
Furthermore, by definition the link variable is discrete and the regions are represented by a vector of real-valued features $(x_1, x_2, \ldots, x_n)$ extracted from an automatic detection system. Therefore we apply logistic regression to compute $P(L_{ij} = \ell_{ij}|A_i, B_j)$:

$$P(L_{ij} = \ell_{ij}|A_i, B_j) = \frac{\exp\left(\beta_0^{\ell_{ij}} + \beta_1^{\ell_{ij}} x_1 + \cdots + \beta_{2n}^{\ell_{ij}} x_{2n}\right)}{1 + \exp\left(\beta_0^{\ell_{ij}} + \beta_1^{\ell_{ij}} x_1 + \cdots + \beta_{2n}^{\ell_{ij}} x_{2n}\right)},$$

where β's are the model parameters to be estimated and the index $2n$ is the total number of region features from both views. We note that other estimators such as multilayer neural networks can be also used to define $P(L_{ij} = \ell_{ij}|A_i, B_j)$ but we choose logistic regression as it ensures in a straightforward way that the outputs $P(L_{ij} = \ell_{ij}|A_i, B_j)$ are probabilities.

In our multi-view detection problem the object (organ) contains a number of links where every region in one view is connected to all the regions in the other view. Hence, it is intuitive and straightforward to construct a causal structure where all the links are modelled in parallel (see the first top layer in the network depicted in Figure 3.4(a)). Thus, using the context modelling capabilities of Bayesian networks we consider at once all the information available about the object.

Next we estimate the probabilities $P(C_{A_i} = \textit{true}|\{L_{ij} = \ell_{ij}\}_{j=1}^{N_B})$ and $P(C_{B_j} = \textit{true}|\{L_{ij} = \ell_{ij}\}_{i=1}^{N_A})$ where C_{A_i} (C_{B_j}) is the class of region $A_i(B_j)$, $N_A(N_B)$ is the total number of regions in *View-A* (*View-B*) and $\{L_{ij} = \ell_{ij}\}_{j(i=1)}^{N_B(N_A)}$ denotes the set of all links containing $A_i(B_j)$. Given our link class definition in (3.1), we can easily model these conditional dependencies through a causal model using the logical OR. We refer to this



a) **RegNet** b) **ViewNet**

Figure 3.4: Bayesian network framework for representing the dependencies between multiple views of an object

Bayesian network as `RegNet` (see Figure 3.4(a)).

Recall that our main goal is to optimize classification globally in terms of the whole object (organ). Therefore, we construct a second Bayesian network to combine the computed region probabilities from `RegNet` to obtain the probability of a view being *true*. We use a causal-independence model with the logical OR where the cause nodes C_i are the region probabilities, the intermediate nodes I_i are the region classes and the only leaf node is the view probability. This Bayesian network is depicted in Figure 3.4(b) and we refer to it as `ViewNet`. The whole multi-view model based on `RegNet` and `ViewNet` is called `MV-CAD`.

Finally, we combine the view probabilities obtained from `ViewNet` into a single probabilistic measure for the object (organ) as a whole by using different schemes. The first scheme `MV-CAD-Avg` is straightforward–simply averaging both view probabilities. In another more advanced scheme `MV-CAD-LR`, we take into account the class of the object (*false* or *true*) by using a logistic regression model with the estimated view probabilities as input variables.

3.4 Application to breast cancer detection

As mentioned in the introduction, multi-view analysis plays a crucial role in the breast cancer detection on mammograms. Here, we describe the application of the proposed Bayesian network framework in this domain.

3.4.1 Single-View CAD System

The inputs for our multi-view detection scheme are the regions detected by a single-view CAD system[160] that consists of the following main steps (see Figure 3.5):

1. Segmentation of the mammogram into background area, breast, and for MLO, the pectoral muscle.

2. Initial detection of pixel-based locations of interest. For each location in the breast area a number of features are computed that are related to tumor characteristics such as presence of spicules and focal mass. Based on these features, a neural network (NN) classifier is then employed to compute likelihood for cancer. The locations with a likelihood above certain threshold are selected as locations of interest.

3. Region extraction with dynamic programming using the detected locations as seed points. For each region a number of continuous features are computed based on breast and local area information.

4. Region classification as "normal" and "abnormal" based on the region features. A likelihood for cancer is computed based on supervised learning with a NN and converted into *normality score* (NormSc): the average number of normal regions in a view (image) with the same or higher cancer likelihood. Hence, the lower the normality score the higher the likelihood for cancer.

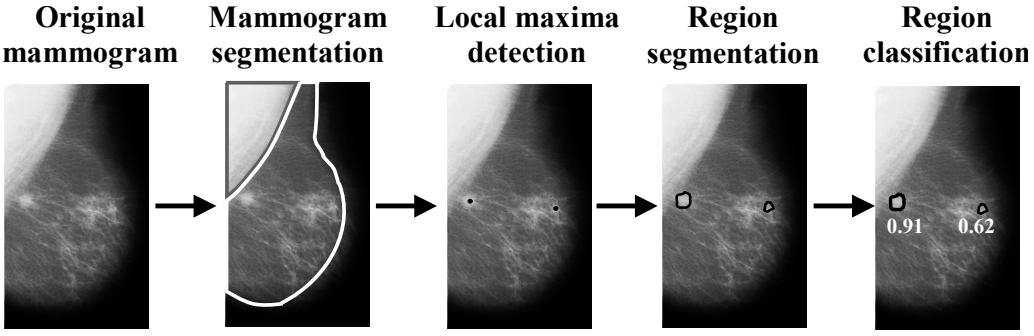


Figure 3.5: Stages in the single-view CAD system

3.4.2 Data Description

The data set we use in this study contain 1063 screening exams (cases) from which 385 were cancerous. The data is a mixture of 795 current cases and 268 prior cases; 33

were cancerous priors with the cancer visible in retrospect. We considered the exams of one patient as independent cases. All exams contained both MLO and CC views. The total number of breasts were 2126 from which 388 had cancer. All cancerous breasts had one visible lesion representing a mass, architectural distortion, or asymmetry in at least one view, which was verified by pathology reports to be malignant (cancerous). Lesion contours were marked by, or under supervision of, an experienced screening radiologist.

For each image (mammogram) we have a number of regions detected by the single-view CAD system. Every region is described by 11 real-valued features automatically computed by the system. These features include the neural network's output from the single-view CAD and lesion characteristics such as spiculation, focal mass, size, contrast, linear texture and location coordinates. Since the only certain information we have about the findings is the one related to the cancer, for each region, based on the ground-truth data, we have a class value of $true$ ("cancerous") if the detected region hits a cancerous finding and $false$ ("normal") otherwise, which may also include hits of benign findings. Every region from MLO view was linked with every region in CC view. For every link we added the binary class values of $false$ ("normal") and $true$ ("cancerous") following the definition in (3.1). We assign analogous binary classes for view, breast and case based on the ground-truth information.

We construct the data such that every row corresponds to one breast observation represented by all feature vectors for the regions in MLO, followed by the regions in CC. The sequence of regions per view was determined by the level of suspiciousness, starting with the most suspicious one. In this study we conduct experiments with two datasets where we select 5 and 3 regions per view with the lowest NormSc. The two datasets are described in Table 3.1. The selection of 5 regions per view leads to data where in only 5 out of 385 cancerous cases there is no TP detected region and thus no $true$ link is available in these cases for training the networks. On the other hand, we have a large number of MLO and CC regions, which are mostly FP. Therefore for the second data set we choose 3 regions per view. However, this leads not only to considerably less FP regions in MLO and CC but also to a higher number of missed TP regions–in total 13 out of 385 cases.

3.4.3 Training and Evaluation

To train and evaluate the proposed multi-view CAD system, we used two-fold cross validation: the dataset is randomly split into two subsets with approximately equal number of observations and proportion of cancerous cases. The data for a whole case belonged to only one of the folds. Each fold is used as a training set and as a test set.

Table 3.1: Description of the two datasets used in the current study

Parameter	Dataset-1 (Data5reg)	Dataset-2 (Data3reg)
Number of regions per view	5	3
Total number of regions (MLO / CC)	10478 / 10343	6358 / 6328
Number of cancerous cases without $true$ links	5	13

At every level (region, view, breast and case) the same data folds were used. Although we use the results from the single-view CAD system, we want to emphasize that the random split for the multi-view CAD system is done independently–the single-view CAD system was trained and tested with ten-fold cross validation on a much larger dataset including regions from cases without CC views.

Bayesian network training. Both `RegNet` and `ViewNet` have been built, trained and tested by using the Bayesian Network Toolbox in Matlab[156]. The learning has been done using the EM algorithm, which is typically used to approximate a probability function given incomplete samples (in our networks the OR-nodes are not observed[168].

Breast data training. As we discussed in the description of our model, we apply two combining schemes–averaging and logistic regression–to compute the probability for a breast being cancerous given the respective view probabilities. For the logistic regression, the input contains view information represented by the probabilities for MLO and CC obtained from `ViewNet` and the minimum NormScs for each view, which are also indicators for view suspiciousness.

Case classification. We compute the likelihood of a case being cancerous based on the computed right and left breast probabilities. The first simplest approach is to take the maximum out of both probabilities. Furthermore, for the `MV-CAD-LR` model, which accounts for the breast classes, we presume that further improvement can be achieved by using the case class. Therefore we perform logistic regression using two inputs: the maximum out of both breast probabilities and the single-view measure for suspiciousness. Thus from the multi-view CAD system, we obtain in total three measures for a case being cancerous: `MV-CAD-Avg-max`, `MV-CAD-LR-max` and `MV-CAD-LR-LR`.

The performance of our multi-view model is compared with that of the single-view CAD system (SV-CAD). For the latter, the likelihood for a view, breast and case being cancerous is computed by taking the likelihood (NormSc) of the most suspicious region. The comparison analysis is done using the Receiver Operating Characteristic (ROC) curve[169] and the Area Under the Curve (AUC), a standard performance measure in the radiologists' practice. The significance of the differences obtained in the AUC measures is tested using the ROCKIT software for fully paired data: for each patient we have a pair of test results corresponding to MV-CAD and SV-CAD[152].

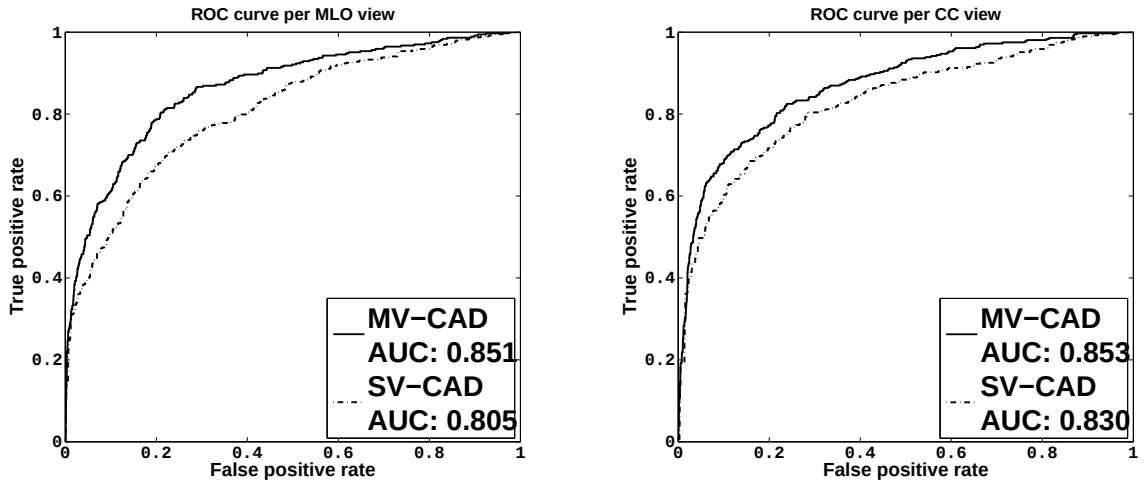
3.4.4 Experiments and Results

Individual view, breast and case classification.

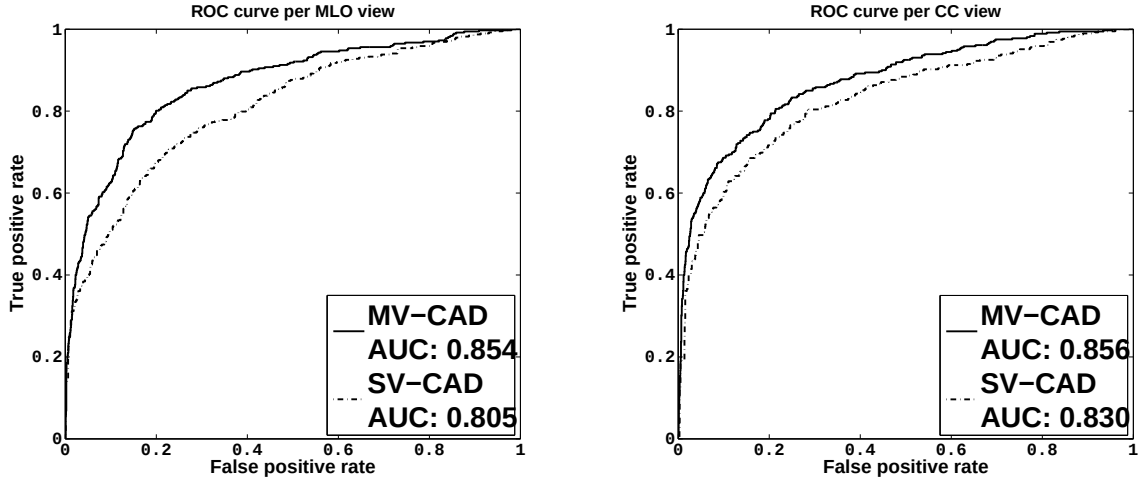
Based on the results from ViewNet, Figures 3.6(a) and 3.6(b) present the classification outcome with the respective AUC measure per MLO and CC view for Data5reg and Data3reg. First we observe that for both MV-CAD and SV-CAD the performance for CC view in terms of AUC is better than that for MLO view. This can be explained by the fact that the classification of CC views is generally easier than that of MLO views due to the breast positioning. At the same time our multi-view system improves considerably upon the single-view CAD system in better distinguishing cancerous from normal MLO views whereas for CC views this improvement is less. Another interesting result is that the largest improvement, especially for MLO view, is observed in the lower scale of the false positive rate (< 0.5).

To check the significance of the difference between the AUC measures we test the hypothesis that the AUC measures are equal against the one-sided alternative hypothesis that the multi-view system yields higher AUC for MLO and CC views. Table 3.2 summarizes the statistical test results by providing the corresponding p-values and 95% confidence intervals of the difference between the AUC measures. The results clearly indicate an overall improvement in the discrimination between cancerous and normal views for both MLO and CC projections. Such an improvement is expected as the classification of each view in our multi-view system takes into account region information not only from the view itself but also from the regions in the other view.

While the view results are very promising from a radiologists' point of view it is more important to look at the breast and case level performance. Tables 3.3 and 3.4 presents the respective AUC (standard error) obtained from MV-CAD and SV-CAD system as well as the one-sided p-values and 95% confidence intervals obtained from the tests on the differences between our multi-view model and the single-view system. Although the simple averaging method MV-CAD-Avg (MV-CAD-Avg-max) tends to show



a) Data5reg



b) Data3reg

Figure 3.6: ROC analysis per MLO and CC view

better distinction between normal and cancerous breasts (cases) with respect to the `SV-CAD`, the difference in the AUC measures is statistically insignificant. However, taking into account the breast classes and performing new training as done in the more advanced `MV-CAD-LR` leads to a significant improvement in the classification outcome. The best performance for both datasets at a case level is achieved for `MV-CAD-LR-LR`, confirming our expectation that further improvement can be achieved by training using the case class. Furthermore, we note that for both datasets, `MV-CAD-LR-LR` yields the same AUCs but with slightly different p-values. To explain this difference we plot the ROC curves; see Figure 3.7. We see that for `Data5reg` improvement in the breast

Table 3.2: AUC (std.error) obtained from the single- and multi-view system per MLO and CC with the respective one-sided p-values and 95% confidence intervals for the difference

View	Method	Data5reg	p-value	Data3reg	p-value
MLO	SV-CAD	0.805 (0.013)	–	0.805 (0.013)	–
	MV-CAD	0.851 (0.011)	0.000 (0.028,0.063)	0.854 (0.011)	0.000 (0.031,0.067)
CC	SV-CAD	0.830 (0.012)	–	0.830 (0.012)	–
	MV-CAD	0.853 (0.011)	0.004 (0.006,0.039)	0.856 (0.011)	0.001 (0.009,0.044)

cancer detection is observed over the whole range of false positive rates whereas for `Data3reg` it is achieved for false positive rates < 0.6.

Table 3.3: AUC (std.error) obtained from the single- and multi-view system at a *breast* level with the respective one-sided p-values and 95% confidence intervals for the difference

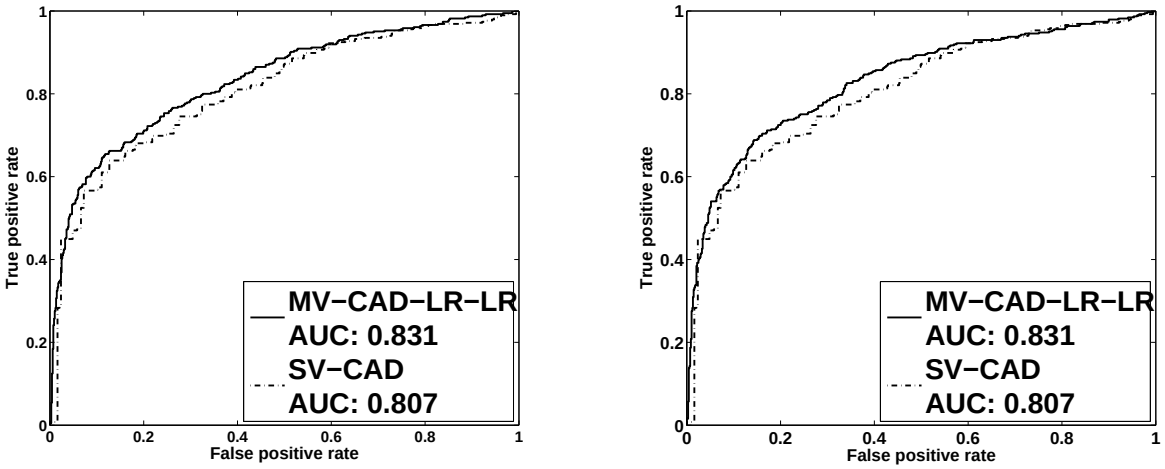
Method	BREAST			
	Data5reg	p-value	Data3reg	p-value
SV-CAD	0.849 (0.012)	–	0.849 (0.012)	–
MV-CAD-Avg	0.862 (0.011)	0.047 (-0.002,0.029)	0.860 (0.011)	0.094 (-0.005,0.026)
MV-CAD-LR	0.868 (0.011)	0.010 (0.003,0.034)	0.865 (0.011)	0.024 (0.000,0.031)

Use of CAD for prescreening of cases.

The results so far presented demonstrate the superior performance of the multi-view system in comparison to its single-view counterpart in terms of individual view, breast and case classification. Here we demonstrate another potential application of the multi-view CAD system to support mammographic decision-making, namely automated prescreening of cases. The objective is to group cases into two basic categories: "suspicious" and "normal" in order to handle these by a different reading protocol.

Table 3.4: AUC (std.error) obtained from the single- and multi-view system at a *case* level with the respective one-sided p-values and 95% confidence intervals for the difference

Method	CASE			
	Data5reg	p-value	Data3reg	p-value
SV-CAD	0.807 (0.014)	–	0.807 (0.014)	–
MV-CAD-Avg-max	0.825 (0.014)	0.040 (-0.002,0.040)	0.819 (0.014)	0.104 (-0.008,0.035)
MV-CAD-LR-max	0.830 (0.013)	0.014 (0.003,0.045)	0.828 (0.014)	0.037 (-0.002,0.041)
MV-CAD-LR-LR	0.831 (0.013)	0.007 (0.005,0.043)	0.831 (0.013)	0.008 (0.004,0.043)



a) Data5reg b) Data3reg

Figure 3.7: ROC analysis per case

One can argue that cases selected as "suspicious" would benefit most from receiving more attention from radiologists. If resources in a screening program only allow for single-reading programs, for example, one might considered a modest extension of the program by double reading only the most suspicious cases. On the other hand, if double reading is practiced and resources are limited, one might consider to use single reading for a subset of cases selected by a CAD system as highly normal, leading to a considerable reduction in the workload. Alternatively, if both radiologists in a double

reading setting do not find an abnormality in a case that is judged highly suspicious by a CAD system, one could present such a case to a third reader performing arbitration, similar to the procedure that is often followed if both readers disagree.

The problem of case prescreening has already been addressed in the literature introducing the concept of using specially trained, non-physician personnel for mammographic prescreening. With the increasing demand for mammography, required training times and shortage of manpower, however, it may be more beneficial to use CAD systems as a prescreening tool. To our knowledge only a few studies discussed so far the application of CAD systems for prescreening of cases[170,171]. The current work contributes to the fund of knowledge in this area by considering the use of the multi-view and the single-view CAD system for the selection of most and least suspicious cases.

The aim of prescreening is optimizing the detection rate while reducing the workload. Because in the breast cancer screening programs the number of normal cases is far larger than that of cancerous cases, the workload is determined by the number of normals to be read. Therefore, for the prescreening task we consider the percentage of detected cancers as a function of a percentage of normal cases with highest likelihood for cancer. Figure 3.8 depicts the results for `MV-CAD-LR-LR` and `SV-CAD`.

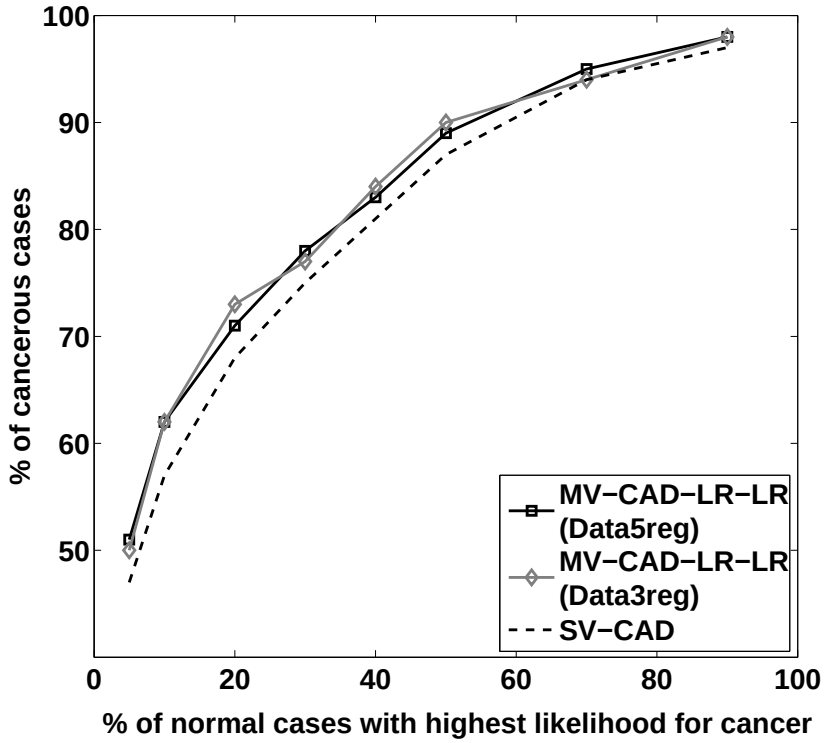


Figure 3.8: Percentages of cancerous cases within subsets of most suspicious normal cases for the single- and multi-view systems

The results demonstrate that using multi-view information leads to overall increase in the number of detected cancers when a subset of normal cases with highest likelihood for cancer is selected. For example, if 10% of the normals are selected then `MV-CAD-LR-LR` on `Data5reg` and `Data3reg` detects 62% and 62% of the cancers against 57% of `SV-CAD`. This trend is especially observed for the lower range ($< 20\%$) of selected normal cases.

When considering the other prescreening task–selection of the least suspicious cases– we would like to minimize the number of cancers missed when a subset of highly normal cases is chosen. In this respect, looking at the upper range of the percentage of selected normal cases in Figure 3.8 (these are the least suspicious cases), we see that the multi-view leads only to a slight reduction in the number of misclassified cancerous cases in comparison to the single-view CAD system. We note that the result that both CAD systems do not detect all the cancers at smaller subsets of normal cases could be explained by the fact that 9% of the cancerous cases included in our study were priors, i.e. cancers that were not detected by the radiologists at the screening stage.

3.5 Conclusions and future research

In this paper we proposed a Bayesian network framework for multi-view mammographic analysis. We showed that the incorporation of expert knowledge in a probabilistic manner led to a higher detection rate of breast cancer compared to a single-view CAD system. This improvement was achieved at a view, breast and case level and it is due to a number of factors. First, we built upon a single-view CAD system that demonstrates a good performance at local breast cancer detection. Second, following the radiologists' practice, we considered multi-view dependencies between MLO and CC views to obtain a single measure for the view, breast and case being cancerous. This was done by: (i) defining links between the regions detected by the single-view CAD system in MLO and CC views (ii) building a probabilistic causal model where all detected regions with their feature vectors and the established region links are considered simultaneously, and (iii) using the logical OR to compute the region and view probability for cancer. Our multi-view scheme also benefits from its Bayesian nature allowing to handle noisy or incomplete information such as the lack of detected or visible lesions in one of the views.

Except the improvement in the individual case-based performance, in the current study we also demonstrated the potential of the multi-view CAD system for prescreening purposes. In contrast to the traditional prompting CAD systems, in this work we considered the problem of breast cancer detection in screening mammography at a

case level. From this perspective, the proposed CAD system could be used to select the most suspicious cases or to group them for batch reading, as a set of difficult cases. In this way, the selected cases would get more attention from radiologists, for example, by providing additional reading. This could help increase the breast cancer detection rate.

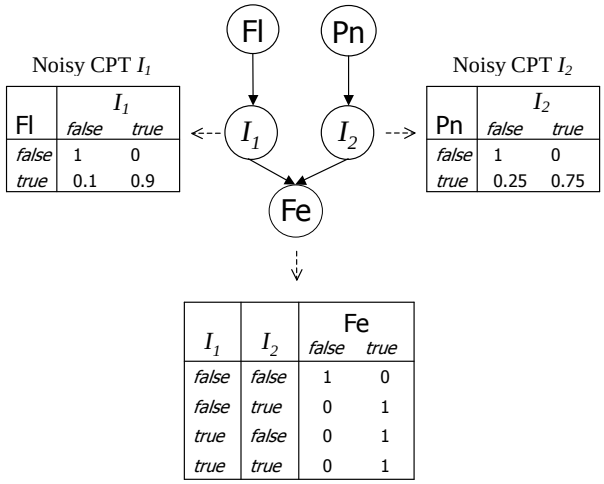
Furthermore, our experiments show that the proposed Bayesian network framework is relatively stable with respect to the number of selected regions per mammogram detected by the single-view CAD. In the current study, we used two versions of the same set of patient cases: one with 5 regions and the other with 3 regions per mammogram. The results indicate that the performance of the models built on both datasets is comparable on individual view, breast and case classification as well as on the selection of most suspicious cases.

Although we demonstrate that the proposed framework has the potential to assist screening radiologists to improve the evaluation of breast cancer cases, we consider a number of directions for extension. First, the current model is based on features that are independently computed per region. However, it is natural to include multi-view features such as the distance to the nipple or correlation features. In such a way, we can explicitly represent multi-view dependencies not only in a qualitative way (through the Bayesian network's structure) but also in a quantitative way (through the input information). This can help improve the system's detection performance. Another possible extension is based on the model structure. Following our Bayesian network framework with using logistic regression and OR-function at a link and view level, we can also apply similar combining schemes at a breast and case level. Thus we can allow for better handling of missing or noisy information in the estimation of the breast/case likelihood for cancer. A third interesting extension of the proposed CAD system is the incorporation of temporal information. In the screening practice, the decision whether a patient has cancer depends not only on the breast multi-view comparison but also on the comparison of current mammograms with previous mammograms of the same patient. The appearance of a new or developing lesion is a strong indication for suspiciousness. Therefore, by integrating multi-view with temporal information in our Bayesian network framework, we can better represent and more accurately model the decision-making process in screening mammography.

Finally, we note that the straightforward nature of the proposed Bayesian network framework allows its relatively easy application to any domain where the goal is computerized multi-view (object) detection.

Appendix

Figure 3.9 depicts an example of a causal-independence model with two cause variables *Flu* (Fl) and *Pneumonia* (Pn) and one effect variable *Fever* (Fe). Probability distributions $P(I_1|\mathrm{Fl})$ and $P(I_2|\mathrm{Pn})$ represent a noise. The interaction function $f(I_1, I_2)$ for the effect *Fever* is the logical OR.



Fl	I_1 false	I_1 true
false	1	0
true	0.1	0.9

Pn	I_2 false	I_2 true
false	1	0
true	0.25	0.75

I_1	I_2	Fe false	Fe true
false	false	1	0
false	true	0	1
true	false	0	1
true	true	0	1

Figure 3.9: Example of a noisy-OR model

Then the probability of having fever given the states of Fl and Pn is computed as follows:

$$\begin{aligned} P(\mathrm{Fe} = true|\mathrm{Fl}, \mathrm{Pn}) &= \sum_{f(I_1,I_2)=true} P(\mathrm{Fe} = true|I_1, I_2)P(I_1|\mathrm{Fl})P(I_2|\mathrm{Pn}) \\ &= P(I_1 = \textit{true}|\mathrm{Fl})P(I_2 = \textit{true}|\mathrm{Pn}) + \\ &\quad P(I_1 = \textit{true}|\mathrm{Fl})P(I_2 = \textit{false}|\mathrm{Pn}) + \\ &\quad P(I_1 = \textit{false}|\mathrm{Fl})P(I_2 = \textit{true}|\mathrm{Pn}). \end{aligned}$$

For example, given the evidence of $\mathrm{Fl} = true$ and $\mathrm{Pn} = true$ then we obtain

$$P(\mathrm{Fe} = true|\mathrm{Fl}, \mathrm{Pn}) = 0.9 \cdot 0.75 + 0.9 \cdot 0.25 + 0.1 \cdot 0.75 = 0.975,$$

indicating the combined influence of both causes on the probability of having fever.

Matching mammographic regions in mediolateral oblique and cranio caudal views

4

Maurice Samulski and Nico Karssemeijer

Original title: Matching mammographic regions in mediolateral oblique and cranio caudal views: A probabilistic approach

Published in: Proceedings of SPIE Medical Imaging 2008: Computer-Aided Diagnosis. Volume 6915, pp. 69151M

Abstract

Most of the current CAD systems detect suspicious mass regions independently in single views. In this paper we present a method to match corresponding regions in mediolateral oblique (MLO) and craniocaudal (CC) mammographic views of the breast. For every possible combination of mass regions in the MLO view and CC view, a number of features are computed, such as the difference in distance of a region to the nipple, a texture similarity measure, the gray scale correlation and the likelihood of malignancy of both regions computed by single-view analysis. In previous research, Linear Discriminant Analysis was used to discriminate between correct and incorrect links. In this paper we investigate if the performance can be improved by employing a statistical method in which four classes are distinguished. These four classes are defined by the combinations of view (MLO/CC) and pathology (TP/FP) labels. We use distance-weighted k-Nearest Neighbor density estimation to estimate the likelihood of a region combination. Next, a correspondence score is calculated as the likelihood that the region combination is a TP-TP link. The method was tested on 412 cases with a malignant lesion visible in at least one of the views. In 82.4% of the cases a correct link could be established between the TP detections in both views. In future work, we will use the framework presented here to develop a context dependent region matching scheme, which takes the number and likelihood of possible alternatives into account. It is expected that more accurate determination of matching probabilities will lead to improved CAD performance.

4.1 Introduction

One of the major challenges in computer-aided detection (CAD) of mammographic masses is the reduction of false positives while sensitivity is maintained. Although many studies report that CAD systems improve the radiologists' accuracy, its effectiveness is not undisputed. In particular, CAD systems that give a high number of false positive markings results in radiologists not to have sufficient confidence in CAD results[172].

Screening usually consists of two-view mammography, i.e., a mediolateral oblique (MLO) and a cranio caudal (CC) film is obtained from both breasts. Using both views in screening improves the chance of detecting abnormalities, mainly due to additional information from the cranio caudal that allows the lesion to be seen more easily on the mediolateral oblique view. Furthermore, it reduces the number of false positives by offering an additional perspective in which superimposition of normal breast structures in the MLO view simulating a suspect lesion can be recognized as such.

One of the most important steps in multi view CAD techniques is to match corresponding regions in the available views. Few studies have been devoted to the investigation of methods for finding corresponding regions in different mammographic views. Paqueralt *et al.*[53] developed a two view approach, by calculating a correspondence score for each possible combination of segmented structures. Combining this correspondence score with the single view detection score resulted in a significant improvement of their detection results. Highnam *et al.*[173] used a compression model to determine a curve in the mediolateral oblique mammogram which corresponds to potential positions of a point in the cranio caudal mammogram. Recently, Wu *et al.*[49] developed a CAD system that incorporates information from two-view mammograms and bilateral mammograms where the corresponding region of interest on the contralateral mammogram is identified using a regional registration technique. Qian *et al.*[162] designed an ipsilateral multiview CAD system where a region of interest in one view is matched with a region of interest in the other view based on their projection distance and analyzed for corresponding shape and characteristic features.

In a previous project, van Engeland *et al.*[160] presented a method to match suspicious regions segmented by a single-view CAD system. For all possible region combinations in the MLO and CC view, a feature vector was calculated containing a number of features that describe the similarity between both regions and the likelihood of malignancy of both regions computed by single-view analysis, whereupon a correspondence measure was being determined using a Linear Discriminant Analysis (LDA) classifier. Finally, for every region in one view, the region in the other view with the highest correspondence score was selected as the corresponding candidate. Using the obtained

correspondences in a multi-view CAD scheme resulted in an significant improvement of the lesion based detection performance[51]. However, in the case based evaluation there was no improvement. The main reason is that links between a TP region in one view and a FP region in an other view tend to lead to more suspicious ratings of the false positives and less suspicious ratings of the true positives, severely degrading the case based detection performance. Even a small number of TP-FP links can negate the improvement of the lesion based performance. Another reason that the improvement of the lesion based performance is not seen in the case based performance is that the lesions of which the mass likelihood was increased after applying the two-view CAD scheme, were already very suspicious in the other view.

In this work we aim to reduce the number of incorrect links and achieve an optimal balance between finding the correct TP-TP links and discarding the TP-FP links. The primary idea is to distinguish four classes rather than two and use a statistical approach. After introducing the region linking method, we will present some preliminary results and discuss future work.

4.2 Materials and methods

4.2.1 Dataset

The digitized mammograms that were used in this study have been obtained from the Dutch breast cancer screening program. In this study, 412 cases with both a MLO and CC view available were used, 41 prior and 371 diagnostic mammograms. A diagnostic mammogram is taken after a sign or symptom of breast cancer has been found, and a prior mammogram is the screening mammogram taken before breast cancer was diagnosed. In each case there was a malignant lesion visible in at least one of the views. Approximately one half (213 cases) was digitized with a Lumisys 85 digitizer, and the other half (199 cases) was digitized with a Canon CFS300 digitizer. All mammograms were digitized at a pixel resolution of 50 μm and averaged down to a resolution of 200 μm, maintaining a gray value depth of 12 bits.

We will introduce first our single view detection scheme which we will briefly explain in section 4.2.2. In section 4.2.2 we will present our region matching procedure.

4.2.2 Single view detection scheme

To each image in the dataset a CAD scheme was applied and consists of the following steps (Figure 4.1):

- Segmentation of the mammogram into breast tissue, pectoral muscle (if image is a MLO view), and background area.

- Initial detection step resulting in an image representing the likelihood of malignancy and a number of suspect image locations (local maxima in the likelihood image).

- Region segmentation, by dynamic programming, using the suspicious locations as seed points.

- Final classification step to improve the prediction of malignancy using region features.

These steps will be described in more detail in the following paragraphs.

Segmentation of the mammogram The first step of our CAD scheme is the segmentation of an image into breast tissue and background, using a skin line detection algorithm. Additionally, it finds the edge of the pectoralis muscle if the image is a MLO view[146]. After these steps, a thickness equalization algorithm is applied to enhance the periphery of the breast[147]. A similar algorithm is used to equalize background intensity in the pectoralis muscle, to avoid problems with detection of masses located on or near the pectoral boundary.

Initial mass detection step In this step we use a multi-scale technique for the detection of stellate patterns, based on a statistical analysis of a map of the texture orientation in mammographic images through the use of operators sensitive to stellate patterns. The same algorithm is used to indentify patterns of radial gradient vectors, rather than radial spiculations. For each pixel inside the breast area this results in a small number of features calculated that represent presence of a mass and the presence of spiculation[44]. A neural network classifies each pixel using these features and assigns a level of suspiciousness to it. The neural network is trained using pixels sampled inside and outside of a representative series of malignant masses. The result is an image in which pixel values represents the likelihood that a malignant mass or architectural distortion is present. This likelihood image is then slightly smoothed and a local maxima detection is performed. A local maximum is detected when the likelihood is above a certain threshold and no other nearby locations have a higher likelihood value. This results in a number of suspicious locations. Finally an algorithm searches for local maxima that are located closer than 8 mm together and remove multiple candidate locations to avoid multiple suspicious locations on the same lesion.

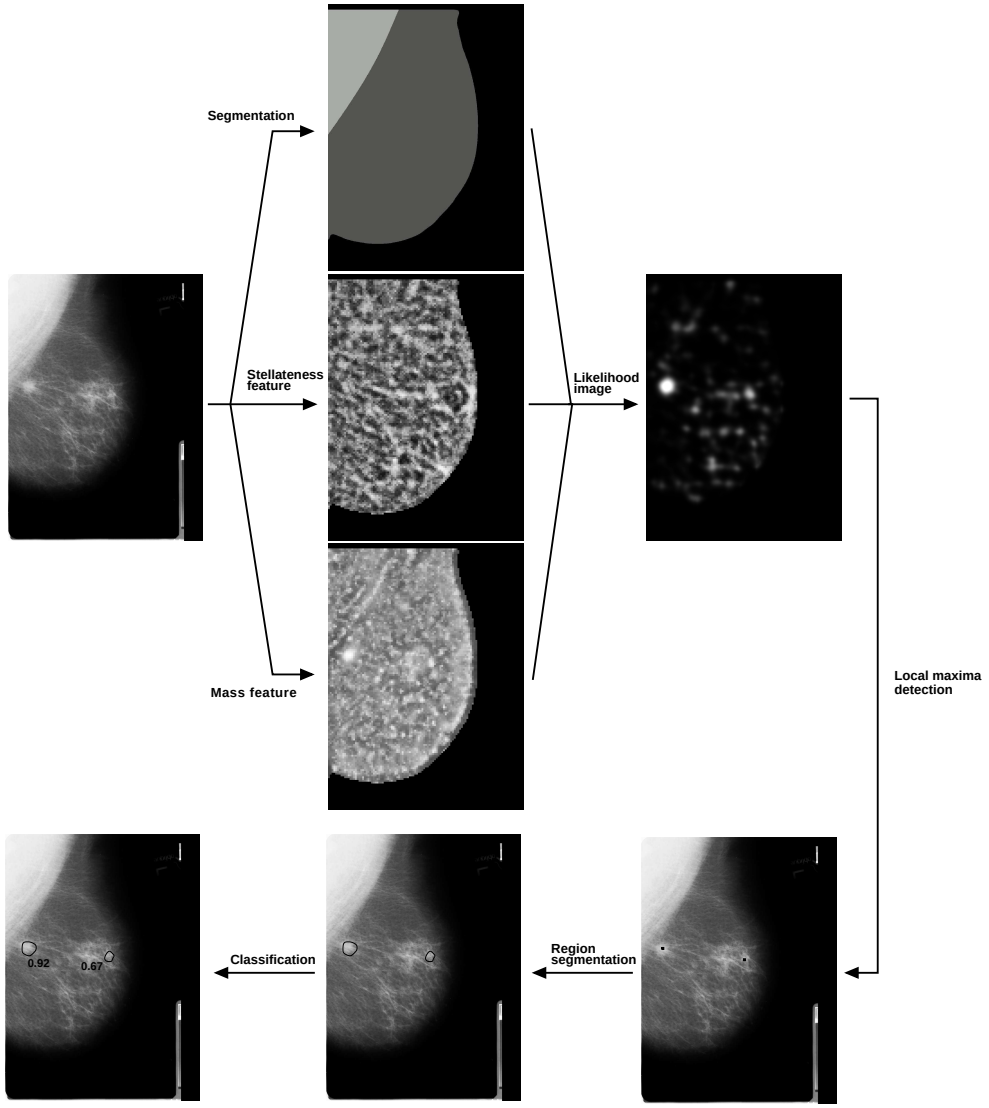


Figure 4.1: Schematic overview of the single view CAD scheme employed in this paper. First the mammogram is segmented into breast tissue, background tissue and the pectoral muscle. We then calculate at each location two stellateness features for the detection of spiculation and two mass features for the detection of a focal mass. A neural network classifier combines these features into a likelihood of a mass at that location, resulting in a likelihood image. The most suspicious locations on the likelihood image (bright spots) are selected and used as seed points for the region segmentation. After that, features are calculated for each segmented region. Finally a second classifier combines these features into a malignancy score that represents the likelihood that the region is malignant.

Region segmentation Each of the detected local maxima in the previous step are used as seed points for region segmentation, based on dynamic programming[148].

Final classification For each segmented region, 81 features are calculated related to lesion size, roughness of the boundary, linear texture, location of the region, contour smoothness, contrast, and other image characteristics. A second neural network combines these features into a malignancy score that represents the likelihood that the region is malignant.

Multi view scheme For every region in one view an annular search area is defined in the other view based on the distance to the nipple, as this is a quite reliable landmark on a mammogram for correlating lesions in MLO and CC views. This reference point is also used by radiologists and remains fairly constant. To define the search area width, we used an annotated database containing 424 cases with a mass lesion that is visible in both the MLO and CC view. For varying width of the search area, the percentage of lesions in the corresponding view that is within this search area is shown in Figure 4.2. A corresponding lesion is within the search area if the difference in radial distance to the nipple is less than half the search area width. The nipple location was estimated by a simple procedure that assumes that the nipple is the point on the skin contour with the largest distance to the chest or pectoral muscle. If we set the search area width to 48 mm, almost all corresponding lesions are within this search area. For large breasts the search area is only a fraction of the breast area but for smaller breasts the search area covered almost the whole breast area. Figure 4.3 shows that there is a small correlation (correlation coefficient 0.17) between breast area and the absolute difference in radial distance between the nipple and the lesion in one view and the radial distance between the nipple and the corresponding lesion in the other view. Based on this, we set the width of the search area from 40 to 48 mm depending on the breast area in order to reduce the number of false positive candidate regions.

4.2.3 Correspondence features

The following paragraphs describe features we use that are invariant to compression and positioning and have high correlation between the values in the MLO and CC view.

Radial distance to the nipple

The radial distance between the lesion and the automatically estimated nipple position remains fairly constant between views (Pearson's linear correlation coefficient was 0.89

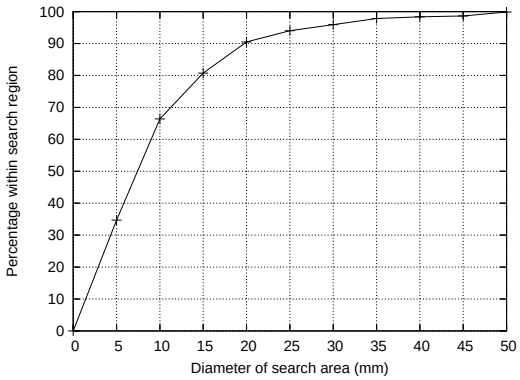


Figure 4.2: Corresponding lesions within the annular search region for varying search area size.

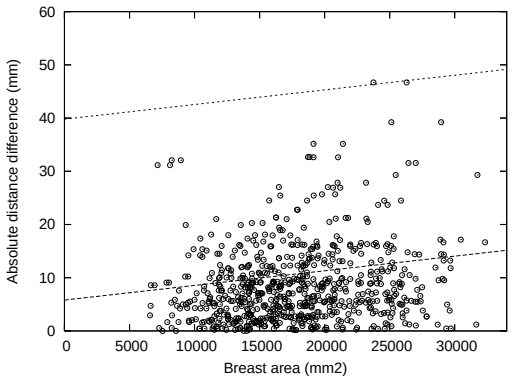


Figure 4.3: Correlation between breast area and the absolute difference in distance between the nipple and the lesion in one view and the distance between the nipple and the corresponding lesion in the other view. Based on this we vary the width of the search area between 40 to 48 mm.

for our annotated database containing 424 cases). The distance feature is defined as follows:

$$distance = \frac{|d_{MLO} - d_{CC}|}{w} \tag{4.1}$$

where d_{MLO} is the radial distance between the lesion and the nipple in the MLO view, d_{CC} is the radial distance between the lesion and the nipple in the CC view and w is the search area width.

Gray scale correlation

We compute a gray scale correlation feature used by Timp *et al.*[55] and Sanjay-Gopal *et al.*[57] for the registration of lesions in temporal mammograms. It is based on the pixel correlation between a region of interest in the current mammogram and a candidate region in the prior mammogram. This region of interest consists all the pixels inside the contour of the region which represents the underlying mass lesion and a band of pixels around the contour which represents the surrounding tissue. This is illustrated in Figure 4.4.

The problem of finding corresponding regions in MLO and CC views is different, since the breast is an elastically deformable soft-tissue structure that is compressed to different extents and in different directions for the two views. To compensate for these effects a modified measure is deduced. First a polar coordinate transformation is

applied to the regions using the center of mass of the regions as the center.

Then Pearson's correlation measure in polar space is calculated, allowing also a rotation ϕ of the CC region with respect to the MLO region. The maximum gray scale correlation over all angles is then used as a feature:

$$gray\_scale\_corr = \frac{\sum_{x,y}(g_{mlo}(x,y) - \overline{g_{mlo}})(g_{cc}(x,y) - \overline{g_{cc}})}{\sqrt{(\sum_{x,y}(g_{mlo}(x,y) - \overline{g_{mlo}})^2)(\sum_{x,y}(g_{cc}(x,y) - \overline{g_{cc}})^2)}} \quad (4.2)$$

$$polar\_corr = max(gray\_scale\_corr(\phi)). \quad (4.3)$$

Entropy

The Shannon's entropy is a measure of the average information carried in a pattern, which is widely used to quantify the smoothness of image texture. Tourassi *et al.*[91] used it for content based retrieval and detection of masses in screening mammograms. We use the observation that Shannon's entropy will be relatively low in homogeneous patterns and increases in inhomogeneous regions to match corresponding regions. We use the absolute difference between the entropy of the polar representation of the region of interest and the candidate region in the other view as a feature.

Histogram correlation

The histogram correlation[160] between the region in the MLO view and the CC view is determined by using two templates. The first template contains pixels inside the region and the second template contains pixels in the band outside the contour as shown in Figure 4.4. A lookup table of gray level values is constructed from the cumulative distribution functions of both the MLO region and candidate region. This lookup table is then used to approximately map the pixel values of the MLO region to pixel values of the candidate region to correct to some degree for differences in exposure. After the application of the lookup table, the gray value histograms for both regions are obtained. The histogram correlation is calculated as follows:

$$HC = 1 - \frac{1}{2}\sum_{g}\left|\frac{H_{MLO}[g]}{T_{MLO}} - \frac{H_{CC}[g]}{T_{CC}}\right| \quad (4.4)$$

where g is the gray level value, H_{MLO} is the gray level histogram of the region in the MLO view, H_{CC} is the histogram of the candidate region in the CC view, and T_{MLO} and T_{CC} are the total number of counts in the MLO, respectively, the CC histogram. The histogram correlations of the inner and outer template are combined into one feature as follows:

$$histogram\_correlation = \frac{HC_{MLO} + HC_{CC}}{2}. \quad (4.5)$$

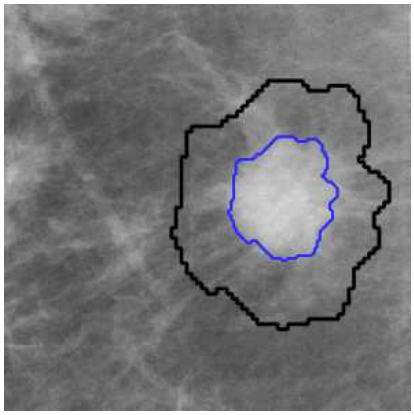


Figure 4.4: Illustration of the inner and outer template used for the computation of the gray scale correlation and histogram correlation features.

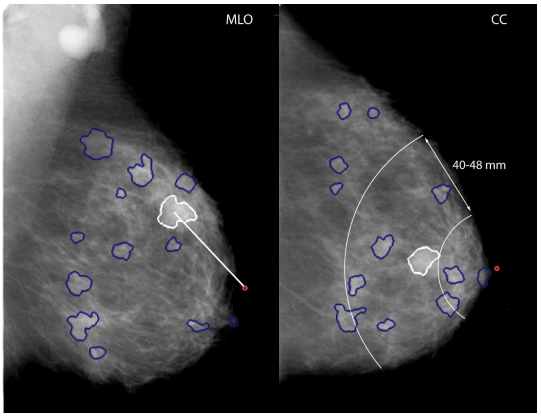


Figure 4.5: Schematic of the search area in the CC view, based on the lesion to nipple distance in the MLO view. The light dot indicates the estimated nipple location.

Mass likelihood

The output of the second neural network classifier in our single-view CAD scheme represents the mass likelihood of a region. From this mass likelihood we derive three features: the mass likelihood of the region in the MLO view, the mass likelihood of the candidate region in the CC view and the absolute difference between the mass likelihood of both views.

Compactness difference

Compactness represents the roughness of an objects boundary relative to its area. This feature is included in the single-view CAD scheme because benign masses often have a round or oval shape compared to a more irregular shape of malignant masses. We use the difference between the compactness of the region in the MLO view and the CC view as correlation feature.

Linear texture difference

The linear texture feature that originates from the single-view CAD system represents presence of linear structures inside the segmented region, as normal breast tissue often has different texture characteristics than tumor tissue. Again we use the difference of this feature between the region in the MLO and CC view as correlation feature.

4.2.4 Classification of region combinations

We can distinguish four classes of region combinations. The first class contains links between TP regions in both views. The second class and third class represent TP-FP combinations and FP-TP combinations, in the MLO and CC views, respectively. The fourth class deals with FP-FP combinations, which primarily include links between normal breast structures for which no ground truth is available. This approach is markedly different from previous research[160] of our group, where a LDA classifier was trained only on TP-TP region pairs (correct combinations) and TP-FP region pairs (incorrect combinations). The resulting LDA classifier output, referred to as correspondence score, was used to select the most likely region combination. The region was linked if the correspondence score exceeded a fixed threshold. When linked, a number of features that describe the resemblance between the best corresponding regions and their likelihood of malignancy was used to train a new two-view classifier. Instead, we propose a statistical approach by using distance-weighted k-nearest-neighbor density estimation and use the matching probabilities in the two-view classifier. We want to differentiate between the TP-FP class and FP-TP class, considering that linking a false positive lesion in one view to a true positive lesion in the other view has a minor effect on the case based performance as this slightly increases the likelihood of malignancy of the false positive lesion. This is relatively a small fraction of the false positives since most mammograms in screening are normal. However, if a true positive lesion is linked to a false positive lesion, the likelihood of malignancy of the true positive lesion decreases, resulting in a decrease in sensitivity of the CAD system.

4.2.5 Distance-weighted k-nearest-neighbor

The k-Nearest Neighbor (kNN) technique is a straightforward yet effective method for density estimation and has a long history in the pattern classification field. A well-known estimate of posterior probabilities can be determined as follows[174]: we use a fixed value of K neighbors, and use the training data to find an appropriate value for volume V. To do this, a small sphere is centered on a new sample $\boldsymbol{x}$ at which we want to estimate the unconditional density $p(\boldsymbol{x})$, and grow the sphere until it contains exactly K neighbors, irrespective of their class. The posterior probability of class C given a point $\boldsymbol{x}$ is given by

$$p(C|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|C)p(C)}{p(\boldsymbol{x})} = \frac{\frac{N_C}{N}\frac{K_C}{N_C V}}{\frac{K}{NV}} = \frac{K_C}{K}$$

where K is the number of neighbors, K_C is the number of neighbors from class C, V is the volume of the sphere, N_C is the number of points in the training dataset from class

C, and N is the total points in the training set. We selected K by employing 10-fold cross-validation on the training set.

This estimation procedure treats each neighbor equally. We use a variant of the typical kNN algorithm proposed by Dudani[175], the distance-weighted k-nearest neighbor method, suggesting that training samples closest to the test sample should be given greater weight than more distant training samples. This is especially more true in posterior probability estimation than in classification, because a larger K is needed for posterior probability estimation as it lessens the effect of discretization. We have defined the neighbor's weight to be the inverse of its distance to the query point.

4.3 Evaluation

To evaluate the performance of our statistical framework, we will compare it to the region linking method used in a previous study[160] by calculating a correspondence score. The correspondence score is defined as the linear combination:

$$\textit{Correspondence score} = L_{TPTP} - \alpha\, L_{TPFP}$$

where L_{TPTP} is the likelihood that the region combination is a TP-TP, L_{TPFP} is the likelihood that the region combination is a TP-FP, and α is a parameter that can be used to tune between the number of TP-TP links that will be established and the number of TP-FP links. When a threshold is put on whether to link two regions or not, increasing α will reduce the number of TP-FP links, but increase the number of missed TP-TP links. We will set α to 0, for the comparison with the linking method from our previous study, where the correspondence score was the output of the LDA classifier. The LDA classifier is trained using only the region combinations, where the region in the MLO and/or CC view is a lesion. The kNN classifier is trained on FP-FP combinations as well.

The classifiers are tested using 2-fold cross-validation, resulting in a correspondence score for every region combination:

$$\begin{bmatrix} C_{0,0} & C_{0,1} & \cdots & C_{0,CC} \\ C_{1,0} & C_{1,1} & & C_{1,CC} \\ \vdots & & \ddots & \\ C_{MLO,0} & C_{MLO,1} & & C_{MLO,CC} \end{bmatrix}$$

Based on this correspondence score matrix, the best corresponding link is selected, that is, the one with the highest correspondence score. For each view we only take, at most, eight candidate regions into account that have the highest mass likelihood. As

Table 4.1: For every view containing a lesion we tested whether this lesion was correctly linked with the lesion in the corresponding view (TP-TP) or linked to a false positive region (TP-FP) using the LDA classifier from previous research and the kNN classifier used in this study.

Method	TP-TP	TP-FP
2-class LDA classifier	694 (80.8%)	164 (19.2%)
4-class kNN classifier	707 (82.4%)	151 (17.6%)

evaluation measure we count the TP detections that were correctly linked to the TP detection in the other view, and the number of TP regions linked with a FP region in the other view. FP-FP combinations were not taking into account.

4.4 Results

Results in Table 4.1 show the number of correct TP-TP links and the number of incorrect TP-FP links using the LDA classifier from previous research[51] and the kNN classifier used in this study. Some improvement was noted in classifying corresponding links but did not reach statistical significance using the McNemar's chi-square test ($p \leq 0.13$).

Additionally, we can apply a threshold to the correspondence score, such that we only establish a link when the score of the best corresponding link exceeds a certain value. The reduction of TP-FP links is important as previous research[51] showed that links between TP and FP regions severely degrade detection performance. When using a fixed threshold of 0.5, the number of TP-FP combinations decreased from 151 to 107, at the cost of TP-TP combinations found which decreased from 707 to 643 for the kNN classifier. Using the same threshold on the correspondence score of the LDA classifier results in a decrease of TP-FP combinations from 164 to 152 at a cost of 14 unlinked TP-TP combinations. This is shown in Table 4.2. If the threshold is chosen such that the percentage of correct TP-TP combinations is 70%, the number of TP-FP combinations decreases with 35 in favor of the kNN classifier. This decrease is significant using the two tailed Fisher's exact test ($p \leq 0.0008$).

Changing the α to 0.3 in the correspondence score formula (Eq. 4.3) for the kNN classifier while maintaining a fixed treshold of 0.5 on the correspondence score, the number of TP-TP combinations decreases from 643 to 587, and the number of TP-FP combinations decreases from 107 to 88.

Table 4.2: For every view containing a lesion we linked the lesion with the best corresponding lesion in the other view, only if the correspondence score is above the treshold. Then we tested whether the lesion was correctly linked with the lesion in the corresponding view (TP-TP) or linked to a false positive region (TP-FP). The region combinations with a correspondence score less than the threshold are shown in the third column. The first two rows of the table show the performance when using a fixed threshold of 0.5. The next two rows show the results when the threshold is chosen such that we achieve 70% correct TP-TP combinations.

Method	**Linked TP-TP**	**Linked TP-FP**	**Unlinked**
2-class LDA classifier with threshold 0.5	680	152	26
4-class kNN classifier with threshold 0.5	643	107	108
2-class LDA classifier with threshold, TP-TP = 70%	604	89	165
4-class kNN classifier with threshold, TP-TP = 70%	604	54	200

4.5 Conclusions and future work

We have developed a two view region matching method to link mammographic regions in MLO and CC views. This method uses the distance-weighted kNN technique to discriminate region links into the four possible categories: TP-TP, TP-FP, FP-TP, and FP-FP. The correspondence score was set to the likelihood that the region combination is a TP-TP link. For every view that contained a lesion we tested whether this lesion was correctly linked with the lesion in the corresponding view. For 82.4% of the TP regions, a correct link could be established, as is shown in Table 4.1. The maximum performance that could have been achieved was 92.4%, because not all cases contained an annotated TP region in both views. Additionally, we can apply a threshold to the correspondence score, such that we only establish a link when the score of the best corresponding link exceeds a certain value. This reduces the amount of TP-FP combinations considerably, especially in 7.6% of the TP regions where there was no annotated TP region in the other view. Changing the α in the correspondence score formula can be used to further reduce the number of false TP-FP combinations, but at the cost of TP-TP combinations found. Previous research[51] showed that links between TP and FP regions severely degrade detection performance. We expect that the decrease in TP-TP combinations has a less negative effect on the detection performance of the two-view classifier whereas the regions are independently analyzed by the single-view CAD system. If the threshold is chosen such that the percentage of correct TP-TP combinations is 70%, the number of TP-FP combinations decreases significantly when using the 4-class kNN classifier (Fisher's exact test, $p \leq 0.0008$).

Several reasons could be given as causes of incorrect links. The most important cause is the occurrence of two regions with the same distance to the nipple and a near

similar feature vector. Also incorrect segmentations such as multiple overlapping regions in the same lesion in the first stages of our CAD scheme is a common cause for incorrect links. The linking performance without using a threshold is better using the 4-class kNN classifier in comparison to the 2-class LDA classifier, but the improvement did not reach statistical significance using the McNemar's chi-square test ($p \leq 0.13$). However, the real performance gain is expected in the next stage of our CAD scheme where we will combine the information from our single view CAD system with correspondence information and the four class probabilities from our linking method.

Another application of this method we recently implemented is linking CAD regions on a mammographic workstation where a radiologist points at a region of interest in one view and the CAD system presents the corresponding region in the other view. The effect of presenting CAD results in both views to the radiologists needs to be investigated further. Future work includes developing a context dependent region matching scheme, which will also take the number and likelihood of possible alternatives into account to improve the case based sensitivity of our single view detection system. To reduce the number of incorrect links that are caused by similar regions with almost the same distance to the nipple, we will investigate additional features to improve linking.

Optimizing case-based detection performance in a multi-view CAD system

5

Maurice Samulski and Nico Karssemeijer

Original title: Optimizing case-based detection performance in a multi-view CAD system for mammography

Published in: IEEE Transactions on Medical Imaging 2011;30(4):1001-1009

Abstract

When reading mammograms, radiologists combine information from multiple views to detect abnormalities. Most computer-aided detection (CAD) systems, however, use primitive methods for inclusion of multi-view context or analyze each view independently. In previous research it was found that in mammography lesion-based detection performance of CAD systems can be improved when correspondences between MLO and CC views are taken into account. However, detection at case level detection did not improve. In this paper, we propose a new learning method for multi-view CAD systems, which is aimed at optimizing case-based detection performance. The method builds on a single-view lesion detection system and a correspondence classifier. The latter provides class probabilities for the various types of region pairs and correspondence features. The correspondence classifier output is used to bias the selection of training patterns for a multi-view CAD system. In this way training can be forced to focus on optimization of case-based detection performance. The method is applied to the problem of detecting malignant masses and architectural distortions. Experiments involve 454 mammograms consisting of 4 views with a malignant region visible in at least one of the views. To evaluate performance, 5-fold cross validation and FROC analysis was performed. Bootstrapping was used for statistical analysis. A significant increase of case-based detection performance was found when the proposed method was used. Mean sensitivity increased by 4.7% in the range of 0.01-0.5 false positives per image.

5.1 Introduction

Breast cancer screening is generally based on two-view mammography in which mediolateral oblique (MLO) and a cranio caudal (CC) projections are obtained from both breasts. When reading mammograms, radiologists combine information from all available views. They compare MLO and CC views, look for asymmetry, and evaluate changes with respect to prior mammograms. CAD systems are nowadays widely used in breast cancer screening. These include sensitive algorithms for the detection of masses and clustered microcalcifications. However, due to a higher rate of false positives, CAD systems do not yet match the performance of human readers. Use of multi-view context is a known weakness of current CAD technology. In most systems mammograms acquired from the same patient are processed and analyzed independently. In this study we focus on development of a CAD system for the detection of masses and architectural distortions that utilizes correspondence between MLO and CC views. Radiologists compare the two ipsilateral mammographic views to decide whether or not a suspicious lesion is present. If a suspicious region in one view has certain features in common with a suspicious region in the other view, there is a higher probability that the region is a true lesion. Using both views in screening also improves the chance of detecting abnormalities, because abnormalities may be (partly) obscured in one projection by overlapping glandular tissue. In addition, by offering an additional perspective, superimposition of normal breast structures simulating a suspect lesion can be recognized, which reduces the risk of false positives.

Methods to combine information from ipsilateral mammographic views to improve computer aided detection of masses have been reported by several researchers. As a first step, it is commonly proposed to match suspected mass regions detected in an initial single-view detection stage. Matching of corresponding regions in MLO and CC views has also been studied as a subject by itself[52,160,176]. Typically, supervised classification is used to compute the likelihood that regions in a pair correspond to the same true positive (TP) region, based on location relative to the nipple and region similarity. Once region correspondence scores have been computed, information from the two views can be fused. Paquerault *et al.*[53] proposed to rank the single-view and correspondence scores within each image, and to average them subsequently. A significant improvement of detection performance was obtained by using this scheme. However, in this preliminary study only abnormal cases with a mass visible in both views were used. With normal cases included the ranking scheme would not work so well, as assignment of high rankings in normal cases would lead to many strong false positives. More recent studies involved fusion schemes based on similarity features and neural networks[51,177] and Bayesian networks[178]. In these studies a significant improvement

of lesion based detection performance was obtained, while case-based performance remained similar. This is unfortunate, as in clinical practice case-based sensitivity, which considers a lesion to be detected if it is reported in either one or two views, is considered to be more relevant. There may be some benefit though if CAD detects a region in both views, as studies suggest that readers are more likely to act on CAD prompts if a region is marked in both views. This motivated Zheng *et al.*[179] to develop a multi-view CAD system that aims to maintain the same case-based sensitivity level while increasing the number of masses being detected on both ipsilateral views. It was found that the multi-view approach reduced the false positive (FP) detection rate by 23.7% while maintaining a case-based detection sensitivity of 74.4%.

Results from the previous studies indicate that existing systems for combining information from ipsilateral views are effective in increasing the suspiciousness rating of subtle lesions if they are paired with a more obvious abnormality in the other view. However, it seems that with existing methods the rating of the more obvious true positive in a pair is hardly affected by being paired with its counterpart in the other projection. In addition, when a true positive is linked to a false positive, the rating of the true positive will generally decrease after applying the multi-view scheme. A combination of these effects may explain why case sensitivity is harder to improve than lesion sensitivity. It is assumed though that if performance of the correspondence classifier is high enough one should be able to improve case sensitivity with multi-view analysis, as in contrast to true positives most false positives do not match with regions in other ipsilateral views.

The aim of this study is to use an extended linking method for matching corresponding regions in MLO and CC views to improve mass detection performance of our single-view CAD system. Specifically, we investigate how multi-view analysis can be made more effective for improving case-based performance. To this end we study modification of the learning rules of the CAD system using the class probabilities from our linking method and the inclusion of correspondence features.

5.2 Preliminaries

In this study we use a single-view detection scheme that consists of the following stages: segmentation and preprocessing, initial detection of suspect image locations, region segmentation, and final single-view classification (Figure 5.1). These steps have been described in detail in previous publications[44,145,180]. In the following paragraphs, a brief description will be given.

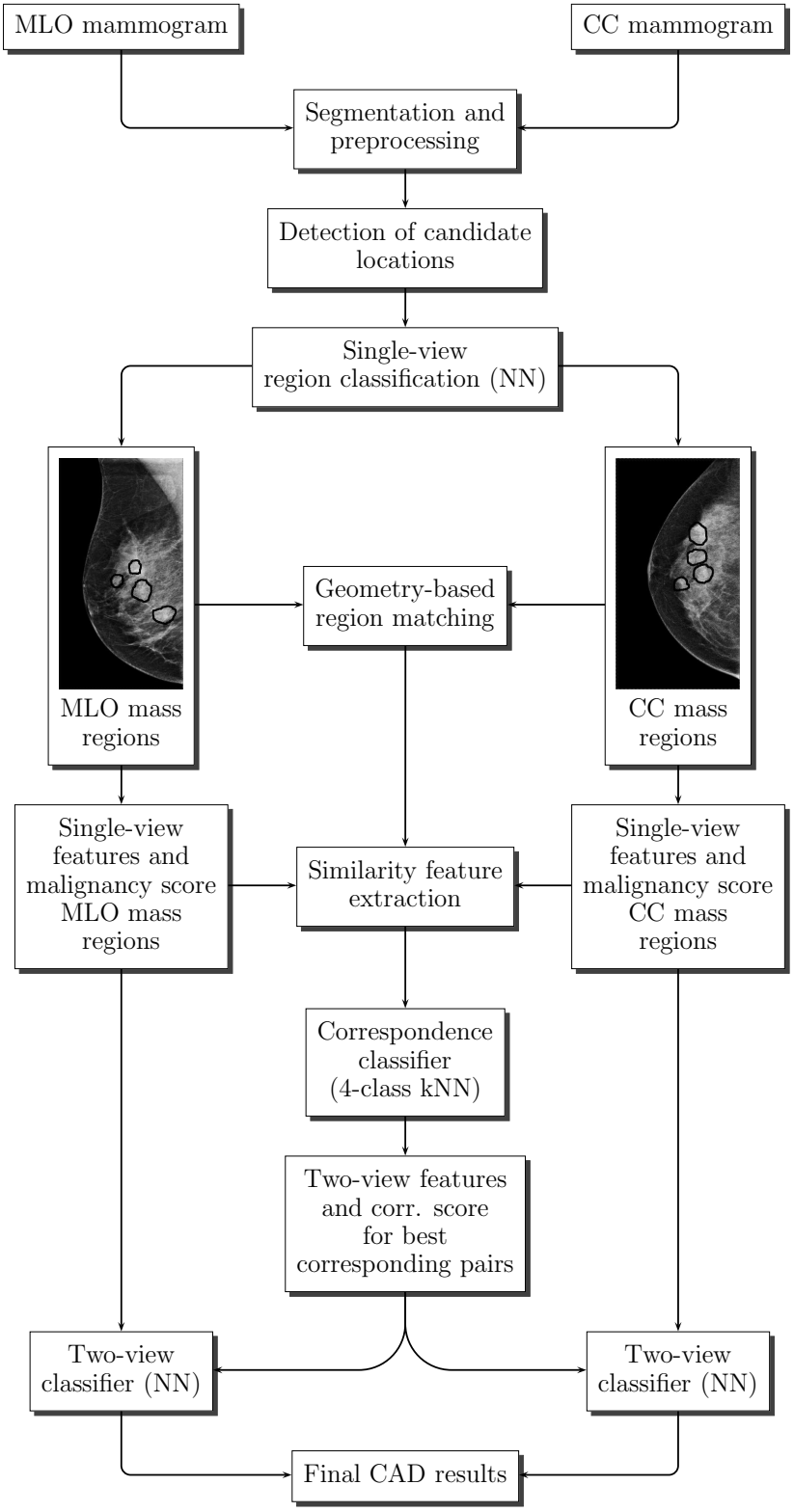


Figure 5.1: Schematic diagram of our two-view computer aided detection system for mass detection on mammograms.

5.2.1 Segmentation and preprocessing

As a first step in the detection scheme, the mammogram is segmented into breast tissue, pectoral muscle (if the image is a MLO view), and the background area. Background pixels are classified by using thresholding in combination with a sequence of morphological operators[146]. Subsequently, the pectoral muscle is segmented from the breast region in two steps. The first is based on straight line estimation using a modified Hough transform as described by Karssemeijer[146]. However, the boundary of the pectoral muscle is generally slightly curved. Therefore, in a second step, the boundary is determined more accurately by an optimal path search near the initial estimate using the dynamic programming method. After these steps, a thickness equalization algorithm is applied to enhance the periphery of the breast[147]. A similar algorithm is used in MLO images to equalize the background intensity in the pectoral muscle, to facilitate detection of masses located on or near the pectoral boundary. Finally, the nipple location was estimated by a procedure that assumes that the nipple is the point on the skin contour with the largest distance to the chest in the CC view or pectoral muscle in the MLO view[51–53,181]. In previous research, this method was evaluated and results appeared be close to the true nipple position (mean error = 12.94mm)[182].

5.2.2 Detection mass regions in single-views

In an initial detection stage, in each image a number of candidate locations are determined which are relevant enough for further analysis given the properties of the pattern in which they are located. To this end, locations in the image area are sampled at an interval of 1.6 mm, and at each location five features are computed that represent the presence of spiculation and a central mass[44,180]. For the detection of stellate patterns, a multi-scale technique is used based on a statistical analysis of a map of the texture orientation in mammographic images through the use of operators that are sensitive to radial patterns of straight lines. The same algorithm is used to identify patterns of radial gradient vectors, rather than radial spiculations, yielding a high response in the central zone of mass lesions. An ensemble of five neural networks, each randomly initialized and trained on a small independent dataset, is used to classify each pixel on a regular grid using these features. Averaging the outputs of the 5 neural networks results in an image in which pixel values represent the likelihood that a mass or architectural distortion is present. This likelihood map is then slightly smoothed and its local maxima are determined. A local maximum is selected as a candidate location when the likelihood is above a certain threshold. This results in a list of locations that are of interest for further investigation. Because the threshold is fixed, the number

of initial locations varies from image to image. As subtle true positive locations may have a low likelihood, a low threshold is chosen to minimize the risk of missing true lesions. This typically leads to a high number of false positive candidate locations. In this study on average 20 locations were selected per image.

In the next stage, each of the detected local maxima in the previous step is used as seed point for region segmentation. We use a method based on polar resampling and dynamic programming which appeared to outperform other methods in a previous study[148]. For each segmented region, region based features are computed representing properties as region contrast, roughness of the boundary, linear texture, relative location in the breast, contour smoothness, lesion size and other region characteristics. In addition to these features, the five central mass and spiculation features, and the mass likelihood score from the initial detection stage are also used by the single-view region classifier.

The neural networks used in our CAD system, are multilayer perceptrons with one hidden layer and an output layer of one neuron. The number of hidden nodes used in the initial pixel classifying stage is five. In the region classifying stage twelve hidden nodes are used. The standard back-propagation (BP) technique with a sigmoid activation function is used to learn the network to map abnormal patterns to a value close to one and normal patterns to a value close to zero. Before training, the features are normalized to zero mean and unit variance using the images in the training set. To avoid overtraining, an independent stopset is used to adaptively determine the number of training cycles.

5.3 Methods

Figure 5.1 shows the schematic outline of the multi-view CAD scheme employed in this paper. The multi-view detection scheme is described in detail below.

5.3.1 Matching regions in ipsilateral views

An important step in multi-view CAD schemes is to match corresponding regions in mediolateral oblique (MLO) and craniocaudal (CC) views of the breast. Because the breast is compressed to reduce the x-ray dose administered to the beast tissue, it is difficult to relate locations of potential mass regions in the MLO view to those in the CC view. The large nonlinear deformation of soft tissue such as the breast makes registration extremely difficult. Most of the matching approaches are therefore based on using a set of landmarks, such as the location of the nipple and points on the pectoral muscle boundary. Two main methods of triangulating a lesion in two projections have been

described in textbooks and journal articles[52,53,79,179,181,183–185]: the arc-based method and the straight-line based method.

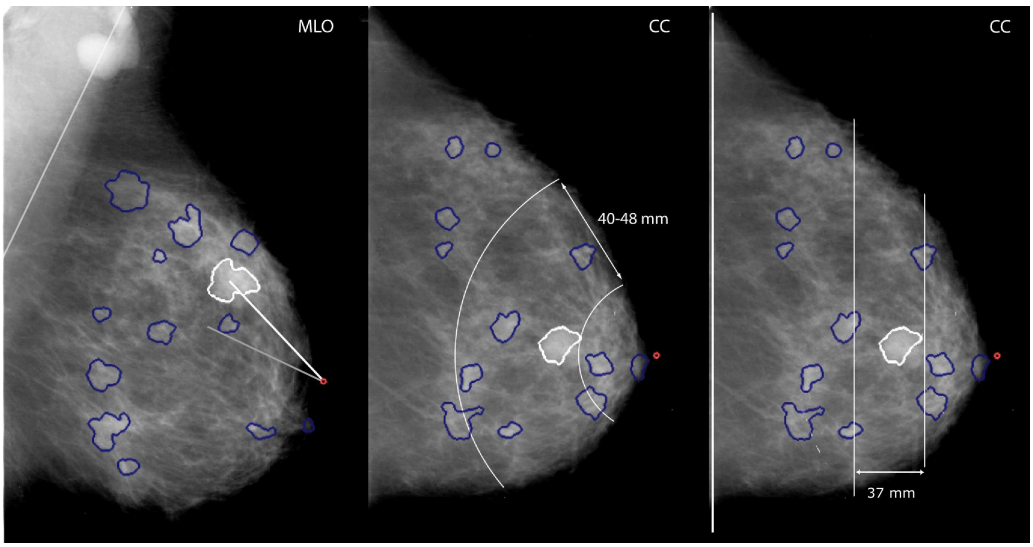


Figure 5.2: Schematic of the geometry-based region matching for finding potential corresponding objects by defining an arc-based search area in the CC view (middle), and a straight-line based search area (right). The light dot indicates the estimated nipple location.

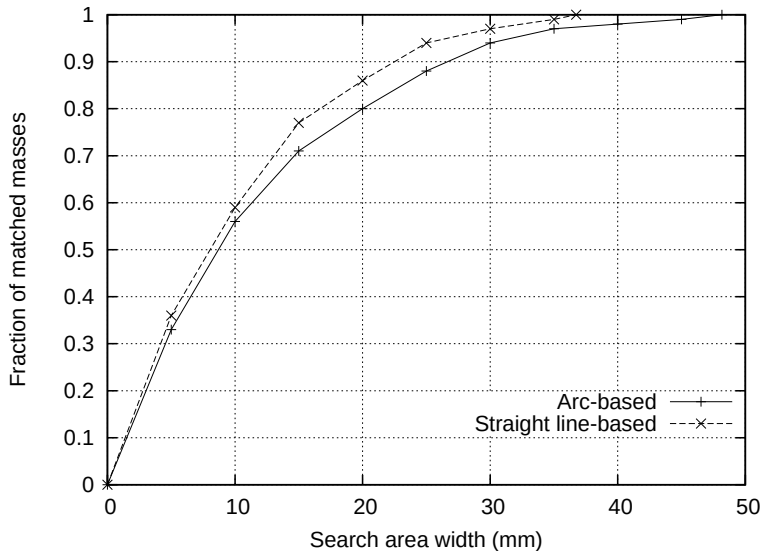


Figure 5.3: Corresponding lesions within the annular search region for varying search area size.

The arc-based method is based on the idea that the distance between the nipple and lesion remains fairly constant during breast compression, which may be explained by the fact that mammographers pull away the breast from the chest wall for optimal positioning. To match a suspicious mass region to a region in the other view, the distance

between the nipple and the center of the mass region is computed. Then an arc is defined in the ipsilateral view with this same distance to the nipple. A search area with a certain width is defined around this arc (Figure 5.2).

The straight line-based method is based on the general concept that the chest wall constrains the deformation during breast compression in such a way that points in the breast move forward with displacement of similar distances under the two views[52,79]. In this method a straight line is defined parallel to the chest wall or pectoral muscle (for the MLO views). In this approach, both the nipple and chest wall has to be detected on both views. Because the chest wall is rarely depicted on the CC view, it was assumed to be parallel to the image edge. The pectoral muscle depicted on the MLO view was automatically detected by the CAD system using a Hough transform based method. To match a suspect mass region in one view to the corresponding mass region in the other view, the distance between the nipple and the center of mass projected onto the centerline is computed. The centerline is the line that is perpendicular to the chest wall and goes through the nipple. This projected distance is then mapped to the centerline of the other view and a straight line parallel to the chest wall is defined. The distance from potential corresponding regions to this straight line is used as a feature to match regions.

To determine the optimal search area width, we used an annotated database containing 424 cases with a mass lesion that is visible in both the MLO and CC view. For varying width of the search area, the percentage of lesions in the corresponding view that was within this search area was determined (Figure 5.3). The arc based method required a search area width of 48 mm to enable matching all pairs. Because there is a correlation between breast area and the absolute difference in radial distance between the nipple and the lesion in the two views, we made the width of the search area dependent on the breast area, in a range 40 to 48 mm. This reduced the number of false positive candidate regions[185]. For the straight line-based method, a smaller search area width of 37 mm was required.

To match lesions in MLO and CC views, for every possible combination of mass regions a number of similarity features are computed. We define these features in such a way that they are insensitive to differences in compression and positioning. The similarity features are described below.

- **Difference in distance**:

$$d_{diff} = \frac{|d_{MLO} - d_{CC}|}{w} \tag{5.1}$$

where d_{MLO} is the distance between the lesion and the nipple in the MLO view, d_{CC} is the distance between the lesion and the nipple in the CC view and w is the

search area width for the arc based-method. For the straight line-based method, d_{MLO} and d_{CC} are the distances from the lesion in the MLO view and CC view, respectively, to the straight line and w is the search area width.

- **Pixelwise correlation**:

  In previous research pixelwise correlation of regions has been successfully used for matching regions in temporal mammogram pairs[55,57]. We apply the same method here to the problem of matching MLO and CC region pairs, even though this method seems naive because of the rotation between the two projections. To compute the correlation a template $\mathcal{T}$ is defined by dilating the segmented region in the source view, i.e. the view in which the region to be matched is located. Pixelwise correlation is defined as

$$PC = \frac{\sum_{r\in\mathcal{T}} (g_s(r) - \overline{g_s})(g_t(T(r)) - \overline{g_t})}{\sqrt{(\sum_{r\in\mathcal{T}} (g_s(r) - \overline{g_s})^2)(\sum_{r\in\mathcal{T}} (g_t(T(r)) - \overline{g_t})^2)}} \tag{5.2}$$

  where r is the location inside the mass template, g_s and g_t are pixel values in the source and target views, and $T(r)$ is the location in the target view corresponding to r, with T a translation to match the centers of mass of the regions. The average pixel values in the mass templates are given by $\overline{g_s}$ and $\overline{g_t}$.

- **Maximum pixelwise correlation in polar space**:

  The problem of finding corresponding regions in MLO and CC views is different from finding corresponding regions in temporal pairs, since the breast is compressed and deformed to different extents and in different directions in the two views. To minimize influence of these acquisition differences on pixelwise correlation, a modified correlation measure is defined. First a polar coordinate transformation is applied to each region using the center of mass of the region as the center. Then Pearson's correlation coefficient is calculated between the transformed regions, using a fixed diameter D of the region, and allowing a rotation ϕ of the CC region with respect to the MLO region. The maximum pixelwise correlation $MPCP$ over all angles is then used as a feature.

- **Entropy**:

  Shannon's entropy is a measure of the average information carried in a pattern, which is widely used to quantify the smoothness of image texture. Tourassi *et al.*[91] used it for content based retrieval and detection of masses in screening mammograms. Entropy will be relatively low in homogeneous patterns and is higher

in inhomogeneous regions. We use the absolute difference between the entropy of the region of interest and the candidate region in the other view as a measure of dissimilarity.

- **Histogram correlation**:

  Histogram correlation[160] can be used as a similarity feature but differences in image acquisition may strongly affect its value. To correct to some degree for differences in acquisition parameters we applied non-parametric histogram matching. In each of the two views a circular area with a radius of 2.5 cm is defined with the center of mass of the region as centerpoint. From both areas the cumulative distribution function is obtained, and a grey level transform is determined which maps the cumulative distribution function of the target area to that of the source area. After applying this transform to the histogram of the target area, differences due to exposure and compression are minimized. It is noted that histograms of the segmented regions will still differ, as the average diameter of true positive lesions is approximately 1.5 cm, which is much smaller than the diameter of the circular areas used for histogram matching. We compute a dual histogram correlation feature DHC. For this purpose two gray value histograms are obtained of both regions using two templates. The first template contains pixels inside the region and the second template contains pixels in a band outside the template. A histogram correlation measure is then calculated for both templates using:

$$HC = 1 - \frac{1}{2}\sum_{g} |H_{MLO}(g) - H_{CC}(g)| \tag{5.3}$$

  where $H(g)$ denotes the fraction of pixels with gray level g in the region template. The histogram correlations of the inner and outer template are combined into a dual histogram correlation feature as follows:

$$DHC = \frac{HC_{inner} + HC_{outer}}{2}. \tag{5.4}$$

- **Mass likelihood**:

  The output of the single-view CAD scheme represents the mass likelihood of a region. From this mass likelihood we derive three features: the mass likelihood of the region in the MLO view, the mass likelihood of the candidate region in the CC view, and the absolute difference between the mass likelihood of both views. When the absolute difference is small between the mass likelihoods of the regions in both views, there is a higher chance that the regions depict the same lesion.

- **Compactness difference**:

  Compactness represents the roughness of an object's boundary relative to its area. This feature is included in the single-view CAD scheme because benign masses often have a round or oval shape while malignant masses generally have more irregular shapes. We use the difference between compactness of regions in the MLO and CC views as a similarity feature. Compactness (C) is computed as the ratio of the squared perimeter (P^2) to the area (A), i.e.,

  $$C = \frac{P^2}{A}$$

  The smallest value of compactness is 4π when the shape is a circle. For more complex shapes, the compactness becomes larger. Therefore this feature is normalized by dividing it by 4π.

- **Linear texture difference**

  A linear texture feature that originates from the single-view CAD system represents the presence of linear structures inside the segmented region[44]. We include this because normal breast tissue often has different texture characteristics than tumor tissue. We use the difference of this feature between the region in the MLO and CC view as a similarity feature.

Using all potential combinations of the mass candidates in the MLO view and the mass candidates in the CC view a k-Nearest Neighbour (kNN) classifier is trained to discriminate MLO/CC region combinations into four possible categories. The first class contains links between true positive (TP) regions in both views, which we will refer to as the TP-TP class. The second class represents combinations between a true positive region and a false positive region in the ipsilateral view (TP-FP), the third class represents combinations between a false positive region and a true positive region in the ipsilateral view (FP-TP). The fourth class encompasses the FP-FP combinations, which primarily include links between normal breast structures for which no ground truth is available. The training and testing procedure is detailed in Section 5.4.2. We define the correspondence score as the likelihood that a region combination represents two true positives. The number of nearest neighbours (k) of the linking classifier was determined experimentally and was set to 27. We chose a high number of k because we use the k-NN classifier for posterior probability estimation rather than hard classification.

In our method each mass candidate detected by the single-view CAD system is linked to the region in the ipsilateral view with the highest correspondence score. As a consequence, by matching a region to the most likely candidate in the ipsilateral view,

the mapping obtained is non-injective. It is noted that in our approach this is not a disadvantage, as we aim at assessing suspiciousness of the individual regions using correlation with the other view as an additional source of information, rather than attempting to obtain a symmetric solution in which region pairs are classified.

5.3.2 Two-view classification

After application of the matching algorithm all regions detected in the single-view scheme are processed further by a new classification module, in which the additional information from the other view is added. For this purpose, a new neural network ensemble with a similar configuration as the single-view classifier is trained. As input a selection of features from the single-view CAD scheme are used, extended with similarity features representing correspondence with the matched region in the ipsilateral view. In addition, the four class probabilities computed by the linking algorithm are included as features. It is noted that the two-view classification results in a new rating of suspiciousness for every individual region, but that we do not attempt to obtain a combined rating for region pairs. In some cases a region cannot be matched with a corresponding region, due to the absence of potential candidate region in the ipsilateral view. The correspondence score needs to exceed a threshold for regions to be matched. When no match is found, we could use the output of the single-view classifier as the final probability of that region. However, this would ignore the fact that the absence of a matching candidate in the other view also gives information about the suspiciousness of a lesion. We know that when a radiologist does not observe a lesion in both views this has influence on interpretation and decision making. Actually, this is a complicated process that involves reasoning about possible causes for not observing the abnormality. In a dense breast, for instance, occlusion of a lesion by glandular tissue is very common. To be able to process all regions with the two-view scheme we use the following approach. When no correspondence can be established for a region (1) the similarity features of that region are set to zero (i.e. there is no correlation between the regions), (2) the class probability that the region is a link between two true positives are set to zero, and (3) the other three class probabilities are set to one third. In this manner, we are able to process unlinked regions with the two-view classifier in the training and testing phases.

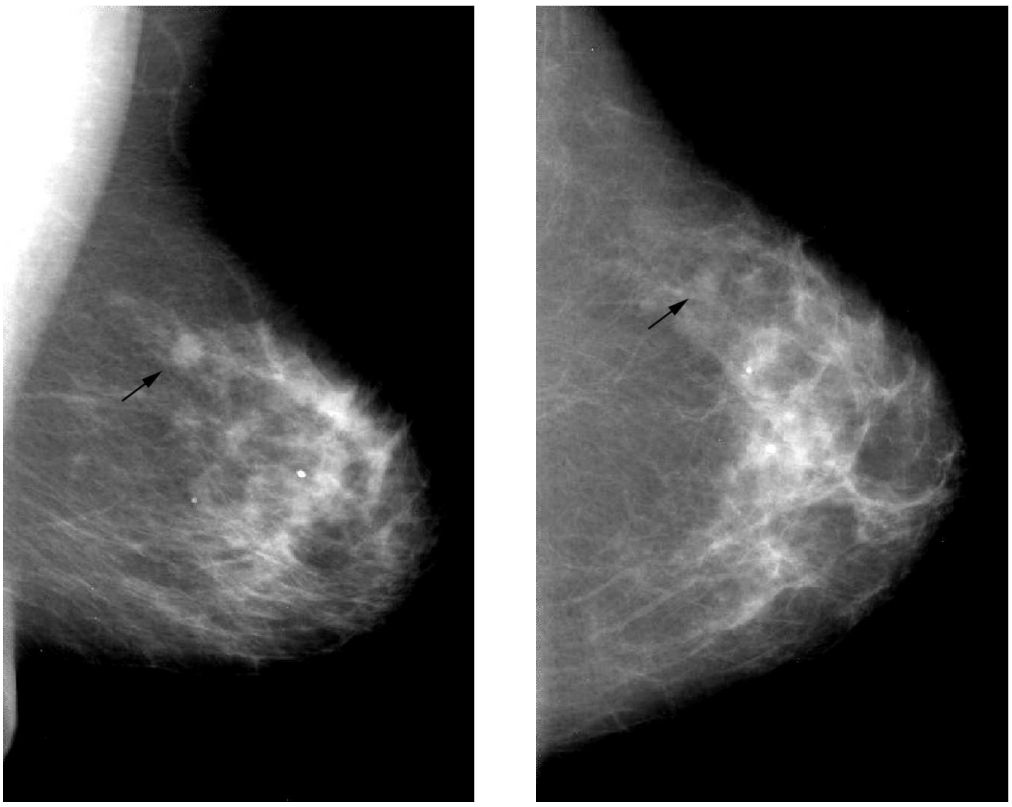
For two-view classification the single-view feature vector was extended with the two-view features and the four class probabilities belonging to this region combination. This resulted in a feature vector of 83 features. To select features for the two-view classifier, we used the sequential floating forward selection (SFFS) algorithm[48], which is based on work by[186] and Spence and Sajda[187]. The cross-validation procedure in

which we select features and evaluate the performance of the two-view classifier is described in detail in the Evaluation section.

5.3.3 Case-based learning

The idea of applying the two-view classifier is that the suspiciousness of a region should be increased when it is linked to a candidate region in the other view with similar characteristics. This is similar to mammographic screening practice: radiologists judge whether or not a lesion is present by comparing both views. If an abnormal region is observed in both views, the likelihood for cancer increases. Suspicious false-positive regions for which no similar suspicious region can be found in the other view will get a lower malignancy score. This is the desired effect we are trying to mimic with two-view analysis. However, linking a true-positive region to a false-positive region will generally result in a lower malignancy score for the true-positive region after applying the two-view classifier. In a previous study, it was shown that this negative effect cancelled out the positive effect the two-view classifier had on true-positive regions that were linked to the correct region in the ipsilateral view.

To improve detection performance on a case level, we propose a method that makes use of correspondence information in training the two view classifier. In this study we use a neural network with the same configuration as the original two-view classifier, but with a modified training scheme that uses the correspondence information to adapt its learning. The idea we explore is to exclude true positive candidate lesions in the training stage if they have weaker signs of malignancy compared to their counterpart in the other view. In this way the two view classifier can be tuned to focus its attention on detecting lesions in at least one of the views, rather than in both views. We expect that with this method case-based performance will increase, possibly at the cost of lesion-based performance. We used two thresholds in this approach to determine when a lesion should be excluded in the training stage. One threshold defines a maximum malignancy score used to determine lesions that are eligible for exclusion. Lesions with higher scores are always included as training pattern. The second threshold defines a minimum difference in malignancy scores between two true positive lesions. The difference between the score of the lesion with the lower rating and that of its corresponding region in the other view should be larger than the threshold in order to make it eligible for exclusion. Both threshold criteria should be met to exclude a pattern from the training set. If the absolute difference between both lesions is small then both patterns are included in the training process. An example of a mammographic case where a poorly visible lesion is excluded from the training process is shown in Figure 5.4.



(a) Left MLO **(b)** Left CC

Figure 5.4: Example of a two-view mammographic case containing an invasive ductal carcinoma marked by the arrow. The mass was detected in the MLO view (a) by the single-view CAD system and was assigned a high malignancy rating. In the CC view (b) the mass was also found but with a very low suspiciousness rating. When using the case-based learning rule, the lesion in the CC view was excluded as primary region in the two-view classifier training.

5.4 Experiments and Results

5.4.1 Materials

The database used in this study consisted of 454 abnormal mammograms (1816 images) and 483 normal mammograms (1932 images) that were verified to remain cancer-free for at least 2 years after the selected exam. All abnormal cases had a visible mass or architectural distortion in at least one view, that was verified by pathology reports to be malignant. All cases consisted of four images: the MLO and CC projections of the right and left breast. For some patients, also prior examinations were available. These were treated as separate cases in this study. Of the abnormal cases, 63 were prior examinations of screen-detected or interval cancers (252 images). 406 cases were digitized with a Lumisys 85 scanner, and 531 were digitized with a Canon CFS300 scanner. All mammograms were digitized at a pixel resolution of 50 μm and averaged down to a

resolution of 200 μm, while maintaining a gray value depth of 12 bits. The malignant masses were annotated by a researcher under supervision of an experienced radiologist. These annotations were used as the ground truth. The initial classifier ensemble, described in paragraph 5.2.2, was trained using a small independent data set (302 images).

5.4.2 Evaluation

Detection performance was tested using free-response receiver operating characteristic (FROC) analysis and 5-fold cross-validation. When splitting the data into a training and test set, we took care that the images belonging to the same case were always assigned to the same set. The cross-validation subsets used in the single-view scheme and in the multi-view scheme were exactly the same, and the test set was used only for testing and never for training in the different stages of the detection scheme. In this way, we ensured that no bias was introduced. For each fold the feature selection is performed on the training set. After the 5-fold cross-validation the malignancy scores for all regions in the 5 test datasets were pooled together and two types of FROC curves were computed, a lesion-based and case-based curve. In the lesion based evaluation sensitivity was computed as the number of lesions detected divided by the total number of lesions. In the case-based evaluation, a case is counted as a true positive when a true positive lesion is detected in at least one of the two views. A CAD region was considered true-positive if the center of mass of the CAD region was inside the annotated region (ground truth).

To get a single performance measure, the mean true positive fraction in a range of false positive rates on a logarithmic scale was computed:

$$MTPF = \frac{1}{ln\,50} \int_{0.01}^{0.5} \frac{TPF(f)}{f} df, \tag{5.5}$$

where f the number of false positives in normal images and $TPF(f)$ is FROC curve, i.e. the true positive fraction as a function of f. We chose the false positive range from 0.01 to 0.5 FP/image as this is the range where radiologists and clinical CAD systems operate in screening practice.

Statistical significance between the single-view and two-view classifier detection performance was determined using the bootstrap method[134,135]. Cases were sampled with replacement from the cross validation set 5000 times. Every bootstrap sample had the same number of cases as the original data set. For each resampling two FROC curves were constructed using the malignancy scores obtained for the two methods to be compared. Next, the difference in $MTPF$ was computed. After resampling 5000

times, it resulted in 5000 values of $\Delta MTPF$. P-values were defined as the fraction of $\Delta MTPF$ values that were negative or zero.

It is noted that in the region matching process not all regions initially detected by the single-view CAD scheme were included, to reduce the computational time. For each view we took at most eight candidate regions into account, where those with the highest likelihood of malignancy were selected.

5.4.3 Results

Results in Table 5.1 show the number of correct TP-TP links and the number of incorrect TP-FP links when no threshold is used on the correspondence score and when we apply a threshold to the correspondence score (probability that a link is a TP-TP), such that we only establish a link when the probability of the best corresponding link exceeds 0.5. With thresholding the percentage of false TP-FP links was 11.1%, the percentage of correct TP-TP combinations was 76.9%, and the percentage of unlinked true positives was 12.0% for the straight-line based method. Based on the results in Table 5.1, we used the straight-line based method and the threshold of 0.5 when training the multi-view classifiers.

Table 5.1: Linking results. For every view containing a lesion we counted the number of times it was correctly linked with the lesion in the corresponding view (TP-TP) and the number of times it was linked to a false-positive region (TP-FP) using the four class kNN classifier. The second and fourth row show the results when a lesion is only linked if the correspondence score exceeds the fixed threshold of 0.5. AB = arc-based method, SLB = straight line-based method.

Linking method	Linked TP-TP	Linked TP-FP	Unlinked
AB without threshold	$724(79.6\%)$	$186(20.4\%)$	–
AB with threshold	$682(74.9\%)$	$104(11.4\%)$	$124(13.6\%)$
SLB without threshold	$729(80.1\%)$	$181(19.9\%)$	–
SLB with threshold	$700(76.9\%)$	$101(11.1\%)$	$109(12.0\%)$

We compared the performance of the two types of multi-view detection schemes and the single-view classifier with each other. The results of the comparisons are listed in Tables 5.2 and 5.3. The second column in Table 5.2 shows the case-based mean true positive fraction $MTPF_{CB}$ obtained, and in Table 5.3 the lesion-based $MTPF_{LB}$. The third and fourth columns show the results of the statistical analysis where the significant differences are shown in bold.

In Table 5.2 the cased-based performances of two types of multi-view classifiers are compared to each other, and to the single-view classifier. Using the case-based learning rule (CBL) in the multi-view classifier resulted in a statistically significant increase in performance compared to the multi-view classifier without using the CBL rule and to the single-view classifier. The difference in case-based performance between the single-view classifier and the multi-view classifier without the case-based learning rule is not statistically significant.

In Figure 5.5 the case-based FROC curves are shown obtained for the single-view classifier, the multi-view classifier without the case-based learning rule and the multi-view classifier with case-based learning. The sensitivity was defined as the fraction of abnormal cases detected as described in the methods section. In Figure 5.6 the lesion-based FROC curves are shown in which the sensitivity was defined as the fraction of abnormal lesions detected.

Table 5.2: Case-based mean sensitivity $MTPF_{CB}$ for the single-view classifier, for the multi-view classifier and the multi-view classifier with the case-based learning rule.

Classifier	$MTPF_{CB}$	Compared to	p-value
Single-view	0.642	-	-
Multi-view	0.658	Single-view	0.0964
Multi-view CBL	0.689	Single-view	**0.0004**
		Multi-view	**0.0198**

Table 5.3: Lesion-based mean sensitivity $MTPF_{LB}$ for the single-view classifier, for the multi-view classifier and the multi-view classifier with the case-based learning rule.

Classifier	$MTPF_{LB}$	Compared to	p-value
Single-view	0.531	-	-
Multi-view	0.573	Single-view	$<$**0.0001**
Multi-view CBL	0.557	Single-view	**0.0180**
		Multi-view	0.1232

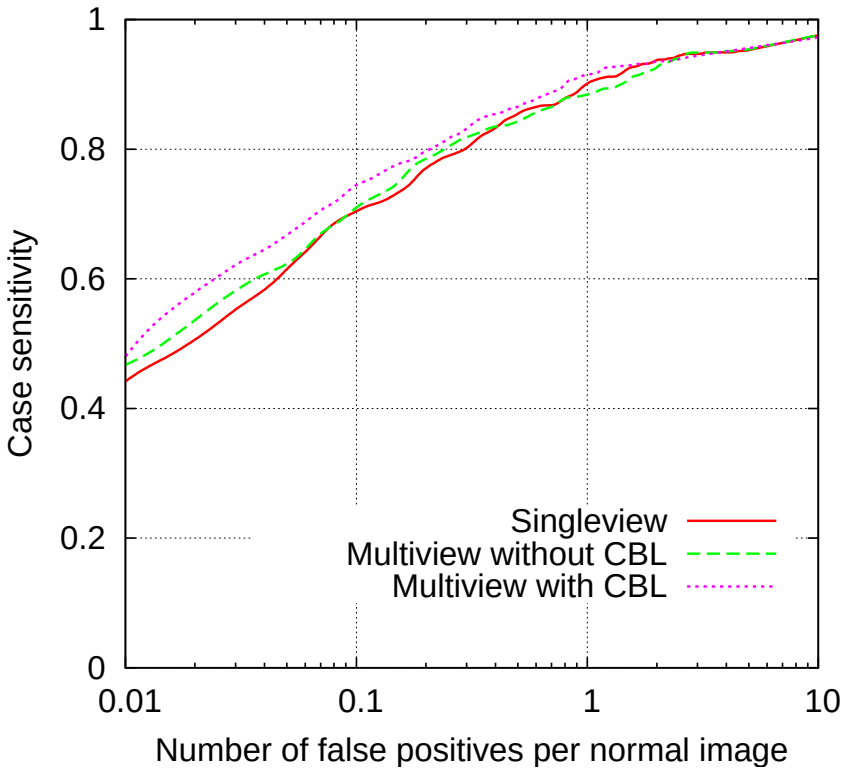


Figure 5.5: Case-based FROC curves for the single-view classifier, the multi-view classifier, and the multi-view classifier with the case-based learning rule.

5.5 Discussion and conclusions

In this paper we investigated a novel method to combine information in a multi-view CAD system. The research was initiated because it was found in previous research that case-based detection performance of a mammography CAD system did not increase with existing methods for combining information obtained from ipsilateral views. In this study, we aimed to improve case-based performance by 1) using an extended region matching scheme using a 4-class region pair classifier, and by 2) introducing a novel learning rule for the two-view detector, in which true positive regions in ipsilateral views are not used as training patterns when they are much less suspicious than corresponding projections in the other view. Results demonstrate that with this new approach case-based detection performance increases significantly in comparison to the single-view CAD system. If appropriate, the case-based learning rule can be adjusted to optimize detecting of abnormalities in both views. The presented scheme can be applied to other CAD applications where multiple projections of lesions are involved.

In a recent paper, Zheng et. al.[52] compared three methods aimed at matching breast

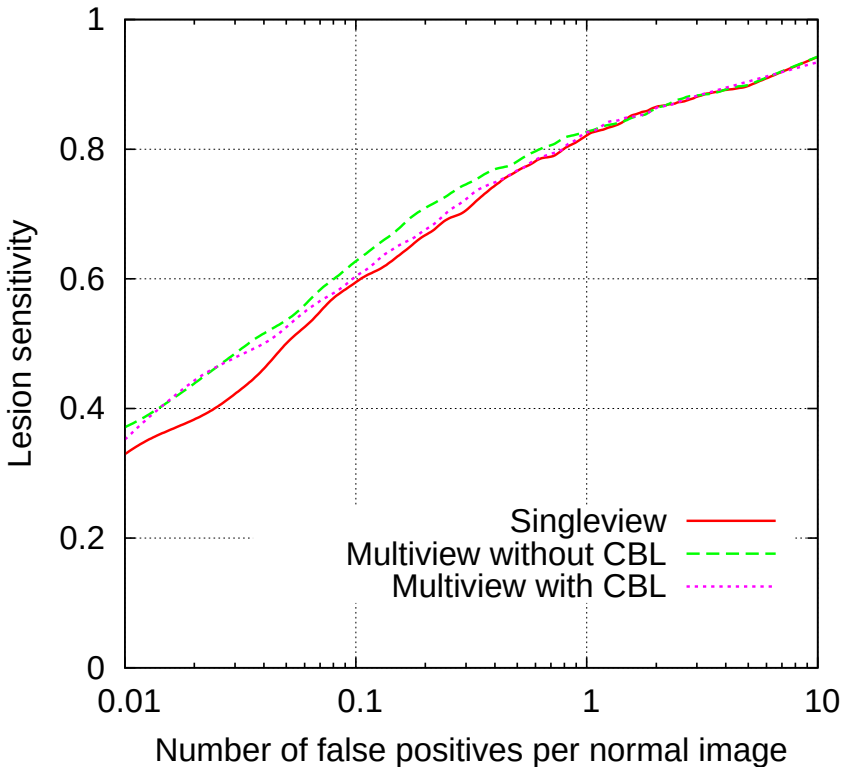


Figure 5.6: Lesion-based FROC curves for the single-view classifier, the multi-view classifier, and the multi-view classifier with the case-based learning rule.

masses depicted on two views and found that the straight line-based method required the smallest search area and achieved the highest level of CAD performance. A preliminary paper of the same group[79] demonstrated that the arc-based method and straight line-based method were comparable in identifying true masses from triangulated observations on two views. However, they favored the arc-based method because only the nipple location is required for localization which made it relatively easy to implement. Based on these results and the evidence that the correlation between distances from the lesion to the nipple in CC and MLO views is high[185], we chose to use the arc-based method in previous studies. In this study we have included both matching methods. Our results confirm those of Zheng et.al. [52] that the straight line-based method requires a smaller search area, but the performance difference is less in our dataset. The assumption that the chest wall can be acceptably represented by a straight line, and the fact that the chest wall is not depicted in most CC views, could affect the performance of the straight line-based method. Our automated nipple estimation algorithm assumes that the nipple is the point on the skin contour with the largest distance to the chest in the CC view or pectoral muscle in the MLO view. Although nipple locations estimated in this way are found to be quite close to the true nipple position, there

are problems with this method when the nipple is not imaged within the field of view, which happens regularly with larger breasts. A better estimation of the nipple position could be made in such situations by extrapolating the nipple position[182]. In the same study it was shown that it can be difficult to locate the nipple in film mammograms by human observers, because the nipple is often hardly visible due to the high optical density of the outer breast edge. Also, there is variability in annotations by humans caused by different methods of annotation. Some annotate the nipple on the skin line, some annotate the nipple at its tip. To investigate if it would be feasible to use the automatically detected nipple positions for correlating lesions in the MLO and the CC view, the authors calculated for two annotated datasets the Pearson correlation coefficient between the distance of the lesion to the nipple in the MLO view and CC view. Neither the automated nipple detection method gives significantly different values of the Pearson correlation coefficient than the manually annotated positions. Hence, it is possible to replace the real nipple locations by the detected positions from the automated nipple estimation algorithm to correlate findings in CC and MLO view. It is noted that full-field digital mammography systems have a larger detector area than the films/screen systems used in this study, and that detection of the nipple using local features is much easier in digital mammography due to the higher dynamic range of digital systems. In a multi-view detection scheme it is important that the number of false correspondences between a TP in one view and a FP in the other view is as low as possible. Links between a TP and FP will generally lead to an increased malignancy rating of the FP and a decreased malignancy rating of the TP, which has a strong negative effect on detection performance, especially when no case-based learning rule is applied. We applied a threshold to the correspondence scores, such that links were only established when the probability of a TP-TP pair was high enough. By doing this the percentage of false TP-FP links decreased considerably, at the cost of losing some correct TP-TP links (see Table 5.1). The negative effect of not linking TP's on the detection performance appeared to be less than negative effect of false TP-FP links. However, to further improve two-view analysis it is important to continue putting effort into improving the linking process.

Using computer aided detection in mammography as a decision support

6

Maurice Samulski, Rianne Hupse, Carla Boetes, Roel Mus, Ard den Heeten and Nico Karssemeijer

Original title: Using Computer Aided Detection in Mammography as a Decision Support

Published in: European Radiology 2010;20(10):2323-2330

Abstract

Objective: To evaluate the effectiveness of an interactive computer-aided detection (CAD) system for reading mammograms to improve decision making.

Methods: A dedicated mammographic workstation has been developed in which readers can probe image locations for the presence of CAD information. If present, CAD findings are displayed with the computed malignancy rating. A reader study was conducted in which four screening radiologists and five non-radiologists participated to study the effect of this system on detection performance. The participants read 120 cases of which 40 cases had a malignant mass that was missed at the original screening. The readers read each mammogram both with and without CAD in separate sessions. Each reader reported localized findings and assigned a malignancy score per finding. Mean sensitivity was computed in an interval of false-positive fractions less than 10%.

Results: Mean sensitivity was 25.1% in the sessions without CAD and 34.8% in the CAD-assisted sessions. The increase in detection performance was significant (p=0.012). Average reading time was 84.7 ± 61.5 seconds/case in the unaided sessions and was not significantly higher when interactive CAD was used (85.9 ± 57.8 seconds/case).

Conclusion: Interactive use of CAD in mammography may be more effective than traditional CAD for improving mass detection without affecting reading time.

6.1 Introduction

Computer aided detection (CAD) was introduced in breast cancer screening as a technology to avoid perceptual oversights and its effectiveness has been demonstrated in many studies [66,69,188]. Nevertheless, there is a continuing debate regarding the usefulness of CAD[71,189]. While most radiologists agree that CAD systems have value because of their high performance in detecting microcalcifications, many believe that current CAD algorithms for masses and architectural distortions have too many false-positives to allow effective use[190–192]. Evidently, more research is needed to improve CAD algorithms. However, the lack of confidence some radiologists have in CAD may also be another reason. In previous research strong evidence was found that the performance of CAD algorithms may not be a problem, but that the concept of CAD may need to be revised[85]. The assumption on which CAD is currently based is that significant lesions initially missed by radiologists will be acted upon when CAD marks them. In practice, however, many lesions are not missed by perceptual oversight but due to incorrect interpretation[5,83,84]. Therefore, it is not surprising that studies reveal that many significant lesions are still missed even when CAD marks them[22,68,193,194]. To prevent such interpretation errors CAD needs to be designed to help radiologists with decision making.

The purpose of this study was to investigate a novel way of using CAD algorithms. In the traditional prompting approach[86,87], CAD results are displayed after the reading is completed, offering the reader a possibility to check if no perceptual failures occurred related to search. In current practice, readers are strongly discouraged to downgrade their findings on the basis of CAD. Compared with the traditional approach, we investigated a method in which CAD marks are only displayed on request during the reading. This novel approach means that when the reader is inspecting a certain region in a mammogram, that particular region can be probed for the presence of any CAD information using a pointer and, if present, only the CAD information about this location is shown. In addition to the CAD mark also the level of suspicion computed by the CAD system is displayed. However image regions deemed normal by the reader are not probed for CAD and thus no other CAD marks elsewhere on the image would be shown. Obviously, this approach will not aid in avoiding perceptual oversights. However, this method has the potential to aid readers in making decisions when they inspect potential lesions, without being distracted by false-positives of CAD.

Our study was motivated by previous research, which demonstrated a significant improvement in detection performance when CAD mass marks were independently combined with reader scores[83]. In that study, CAD marks on regions not reported by the reader were not used, which is similar to the approach investigated here. As

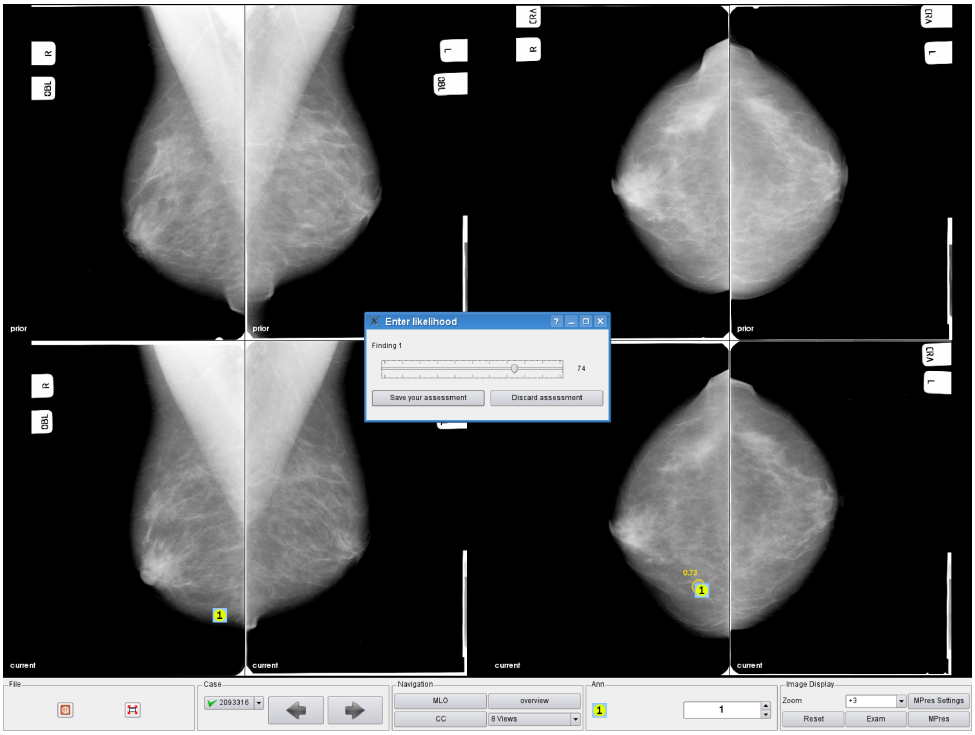


Figure 6.1: The graphical user interface of the CAD workstation used in the observer experiments. The upper row shows prior mammograms and the lower row displays the current screening mammograms that have to be reported. In the case shown here, a reader reported a localized finding in both projections and is asked to assign a malignancy score between 0 and 100 to that finding. In the craniocaudal (CC) view, a CAD region was present at the reported location.

independent combination of reader results with CAD would not be easily accepted in clinical practice, we designed a screening workstation in which readers themselves can combine their interpretation with CAD in an interactive way. To investigate the proposed CAD concept, we conducted a reader study in which 9 readers participated.

6.2 Materials and methods

The institutional review board approved this retrospective study and waived informed consent. For the purpose of this study, a dedicated mammographic workstation was developed that has the basic functionality that screening radiologists expect when they read digital mammograms on electronic displays, including dedicated hanging protocols, zooming, image manipulation, and local contrast enhancement tools. Brightness and contrast were easily adjustable and were set in advance for optimal efficiency. The workstation was equipped with a 30 inch color LCD panel (model FlexScan SX3031W;

Eizo Nanao Technologies Inc., Hakui, Ishikawa, Japan) with a native resolution of 2560 $\times$ 1600. CAD processing is performed on a separate server and results are submitted to the workstation with the image data before a reading session starts. CAD results were obtained from the R2 ImageChecker v8.0 (Hologic, Bedford, MA, USA).

On the workstation (Figure 6.1) the presence of CAD marks can be queried interactively by clicking on suspect regions in the mammogram using a pointing device by the readers. It is not possible to display all available CAD marks at once as in traditional CAD prompting devices. For each queried location, the workstation checks if a CAD mark is available at that location. If a CAD mark is available, it is presented to the reader by displaying the contour of the region detected by CAD along with a computer-estimated malignancy score. The contour of the region is colored based on the malignancy score using a continuous color scale ranging from red to yellow, for respectively high to low malignancy ratings. Previous studies show that giving readers additional information on the likelihood of CAD marks might be helpful in decision making[195–198].

The average number of CAD regions that could be activated was adjustable. Only CAD regions with malignancy ratings exceeding some threshold were included. In the observer study, we adjusted this threshold such that in normal cases the average number of false-positive regions was two per image.

6.2.1 Image database

A total of 120 screening mammograms were selected from the Dutch Breast Cancer Screening program and were digitized using a laser digitizer suitable for medical applications (Lumiscan 85, Lumisys, Sunnyvale, CA, USA) at a pixel resolution of 50 m. The mammograms were averaged down to a resolution of 100 m, maintaining a gray level resolution of 12 bits. From these cases, 40 had a biopsy proven malignant mass, and 80 were cancer-free. Due to the Dutch screening protocol, the majority of the cases had only MLO views available. Of the 120 cases only 25 had additional CC views. All cancer cases selected were subtle cancers that were missed at the original screening and were retrospectively identified as visible. We chose to use cases with missed cancers to maximize the power of our observer experiment. Cases with only microcalcifications were excluded. Each mammogram was presented with the corresponding prior screening mammogram, as is common in screening practice to allow detection of temporal changes. In Table 6.1 the study is summarized.

Table 6.1: Study overview

Total cases	120
Normal cases	80
Cancer cases	40
Cancer cases detected by CAD[a]	33
Available CAD regions[b]	587
Available true-positive CAD regions	41
Available false-positive CAD regions	546

[a] Cancers hit in at least one view by the CAD system at an operating level of 2.0 false-positive markings per image

[b] Regions that could be queried at the operating level of 2.0 false-positives markings per image

6.2.2 Observer study design

Nine readers, of which four were certified screening radiologists and five were non-radiologists with mammogram reading skills, participated in the study. Before the actual observer study, sixty training cases were presented to the non-radiologists. The expert radiologists were presented with fewer training cases due to time constraints. The number of training cases presented to the radiologists ranged from 10 to 30. The training cases served to familiarize the observers with the system, including the reporting functionalities, the interactive CAD functionality, and the controls for adjusting the brightness and contrast.

The observers read the case set in two batches of 60 cases each. Each batch consisted of two sessions. In the first session, 30 mammograms were read with CAD and 30 without. In the second session, CAD was made available for the cases initially read without CAD and vice versa. Each session had a balanced mix of normal and abnormal cases. The order of the cases within each subset was randomized in the two sessions to minimize reading order effects.

The observers were instructed to search for malignant masses and architectural distortions only, and were informed that the study set did not contain microcalcification cases. They were also informed what the approximate proportion of the abnormal cases was. To report abnormalities, readers were asked to mark the finding in the MLO and CC view, and assign a malignancy score on a continuous scale ranging from 0 to 100. Readers were also instructed to mark at least one finding per case, unless a case was so obviously normal that no reasonable finding could be marked. In the

with-CAD session, the readers could query the CAD system by clicking on regions in the mammogram that they were inspecting. Otherwise the reading and reporting was the same as in the non-CAD sessions. They were free to report any finding, regardless if it was marked by CAD or not. There was no limit on the reading time.

6.2.3 Independent combination of readers and CAD

In a previous study the potential contribution of CAD in improvement of mammographic interpretation was investigated by independently combining findings of the readers with detection results of the CAD software[83]. We applied the same method to the experimental data obtained in this study. In that way we could compare the effect of interactive use of CAD during reading with the effect of combining reader reports with CAD independently after the reading is completed. In summary, independent combination was implemented as follows: Only locations in the mammogram that the observers reported were considered. For every finding it was checked whether the location of the finding was marked by CAD and its level of malignancy was determined. If two views were available and the finding was marked in both views, the highest level of malignancy assigned to either of the CAD regions was taken. If the finding was not marked at all by CAD a zero level was assigned. The combined malignancy score of a finding was computed by taking a weighted average of the reader score with the CAD estimated malignancy score.

6.2.4 Statistical analysis

We used localization receiver operating characteristic (LROC) to analyze the data for differences in reader performance between reading with and without using interactive CAD, for individual readers, as well as for the average reader. To determine a LROC, the decision threshold is varied and the correct localization fraction is plotted as a function of the false-positive fraction. The false-positive fraction is defined as the fraction of normal cases recalled as a function of the decision threshold.

For every reader, we determined the cut-off point at which the false positive recall rate was 10%, by thresholding the scores the observer had given to the findings. The primary metric of detection performance was the mean correct localization fraction in the false-positive fraction interval ranging from 0 to 0.1. This interval is chosen because in screening programs radiologists usually have recall rates below 10 percent.

The location of each finding was indicated in the MLO view and CC view. A finding was considered a true-positive (TP), if it had a correct location in at least one of the views. We defined a location to be correct if the distance between the observers'

marked location and the true cancer location was less than 2 cm. The false-positive fraction was estimated from the observers' marked locations in the normal cases. We computed significance of differences between sessions with and without CAD for the average reader using the Wilcoxon signed rank test. Differences with a P value of less than .05 were considered significant. The statistical analysis was performed by using R data analysis software (version 2.9.0; R Foundation for Statistical Computing, Vienna, Austria). The number of times reported and unreported TP and FP CAD regions were queried was computed for every reader. A CAD region was considered queried if the distance between the observers' query location and the centre point of the CAD region was less than 0.5 cm, or if the query location was within the CAD region.

6.2.5 Reading times

Reading times per case were automatically recorded in the reading sessions. Mean reading time per case and its standard deviation was computed for every reader in both reading modes. Reading times exceeding 5 minutes were excluded from the analyses on the basis of the assumption that these excessively long reading times were the result of interruptions during the session. As a result, approximately 3% of all cases were excluded from the time analysis. Average reading times for the unaided session and the session with CAD were calculated. Paired reading times were compared with Wilcoxon signed rank testing. A P value of less than .05 was considered to indicate a statistically significant difference.

6.3 Results

The results of the nine individual readers are shown in Table 6.2. It also shows results obtained by independently combining reader scores with CAD. The mean correct localization fraction of a reader in the false-positive fraction interval ranging from 0 to 0.1 is used as the performance measure. Results show that radiologists did not perform better in this study than the non-radiologists. We computed average LROC curves from all the readers, the non-radiologists, and the radiologists. These are shown in Figure 6.2, 6.3 and 6.4, respectively.

The performance of the average reader increased with CAD at low false-positive rates from 25.1% to 34.8%. Every reader improved their performance using CAD with the exception of reader 8. The difference between reading with and without CAD for the average reader, measured by the performance metric defined above, was statistically significant ($p = 0.012$). Results confirm that performance may also be increased by independent combination with CAD scores, with a smaller increase, however, than

Table 6.2: Reader detection performance in the false-positive fraction interval ranging from 0 to 0.1

	Without CAD TPF10 (%)	With CAD TPF10 (%)	Independent combination TPF10 (%)
Non-radiologists			
1	41.1	51.3	43.3
2	35.3	51.5	41.7
3	16.0	25.9	26.3
4	15.4	25.2	27.4
5	18.3	41.9	26.7
Average	25.2	39.2	33.0
Radiologists			
6	24.3	32.3	33.6
7	24.8	28.8	30.2
8	30.2	25.7	37.0
9	20.2	30.4	30.0
Average	24.9	29.3	32.7
Reader average	25.1	34.8	32.9

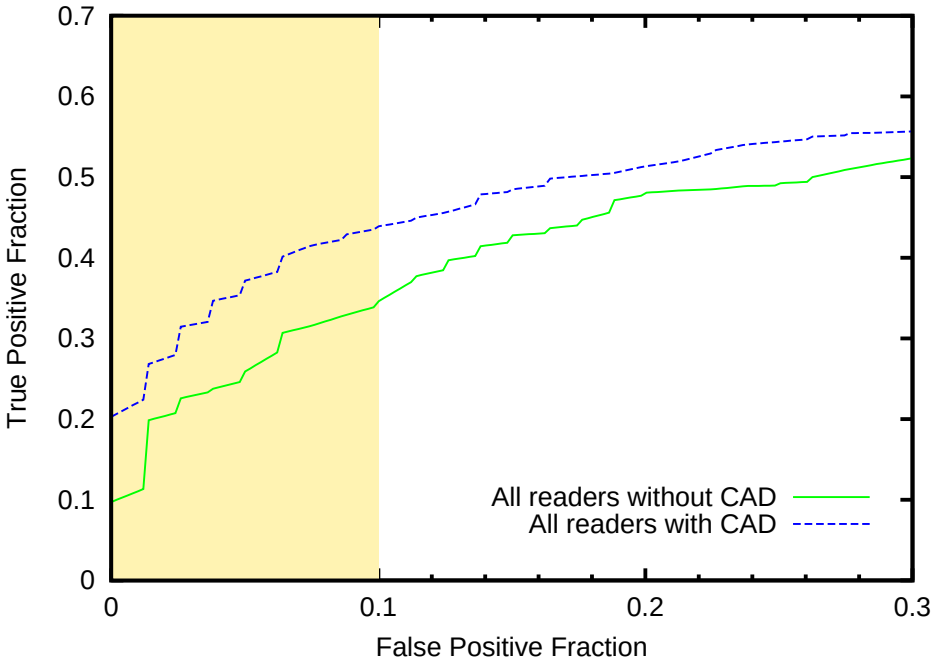


Figure 6.2: Average LROC curves obtained from the nine readers for the detection of cancers with and without using CAD. The false-positive fraction interval ranging from 0 to 0.1, where the mean correct localization fraction is computed, is highlighted in light gray.

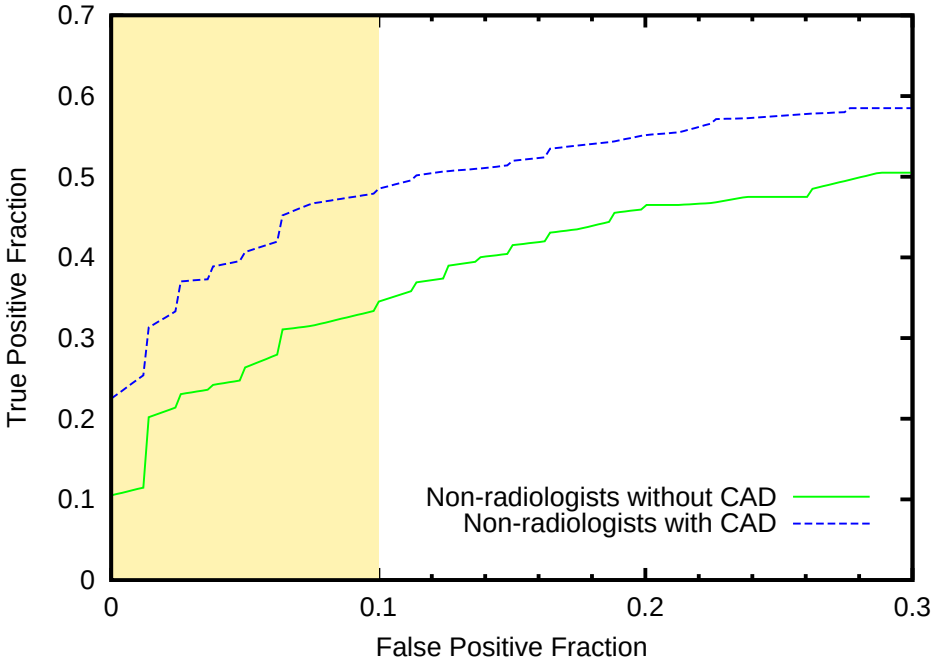


Figure 6.3: Average LROC curves obtained from the five non-radiologists.

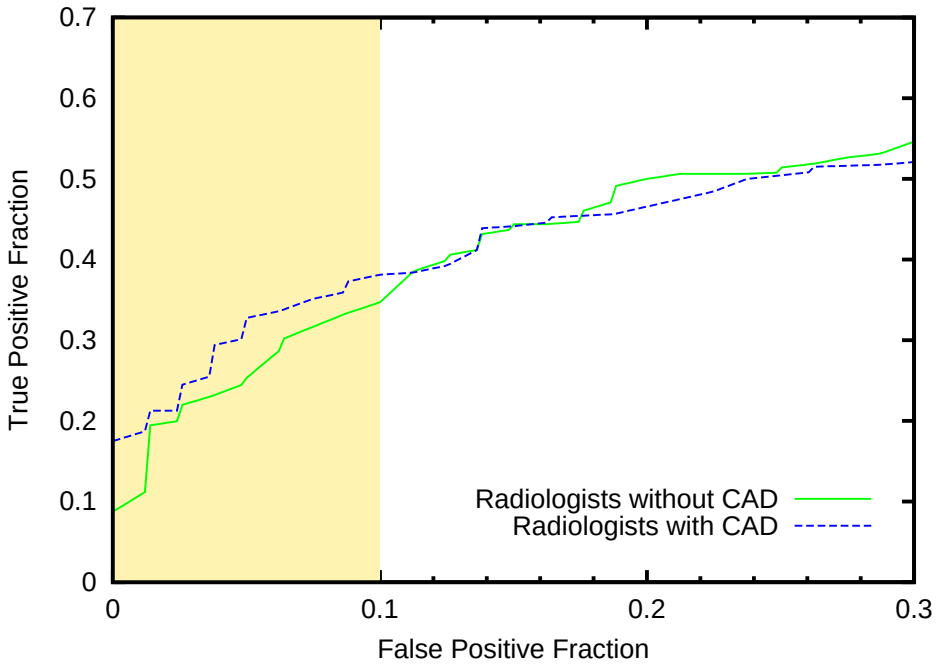


Figure 6.4: Average LROC curves obtained from the four radiologists.

obtained with interactive use of CAD. The difference we found between interactive use of CAD and independent combination is not statistically significant.

As an example, a mammogram of a woman with an invasive ductal carcinoma are shown in Figure 6.5. In this case, 7 of the 9 readers correctly localized the cancer in both sessions, but rated their finding substantially more suspicious in the session with interactive CAD enabled, one reader only located the cancer correctly in the session where CAD was enabled, and one reader did assign a slightly lower rating to the cancer in the session with CAD. In Figure 6.6, the same case is shown with the activated CAD region. The average time to read a case without CAD was 84.7 seconds $\pm$ 61.5. The radiologists read the cases much faster than the non-radiologists. Average reading time in the session with CAD was 85.9 seconds $\pm$ 57.8 per case (Table 6.3). There were no significant differences in reading times for the session with CAD and the session without CAD (p = 0.13) (Table 6.3). The CAD system had a lesion-based sensitivity of 80.4% (41/51) at the operating level of 2.0 false-positive markings per image used in the study. The number of available CAD regions was 587. Table 6.4 shows that on average 274.2 of the 546 false-positive CAD regions (50.2%) were not queried. It also shows that on average 5 of the 41 true-positive CAD regions (12.2%) were not queried. The radiologists queried far fewer false-positive CAD regions than the non-radiologists.

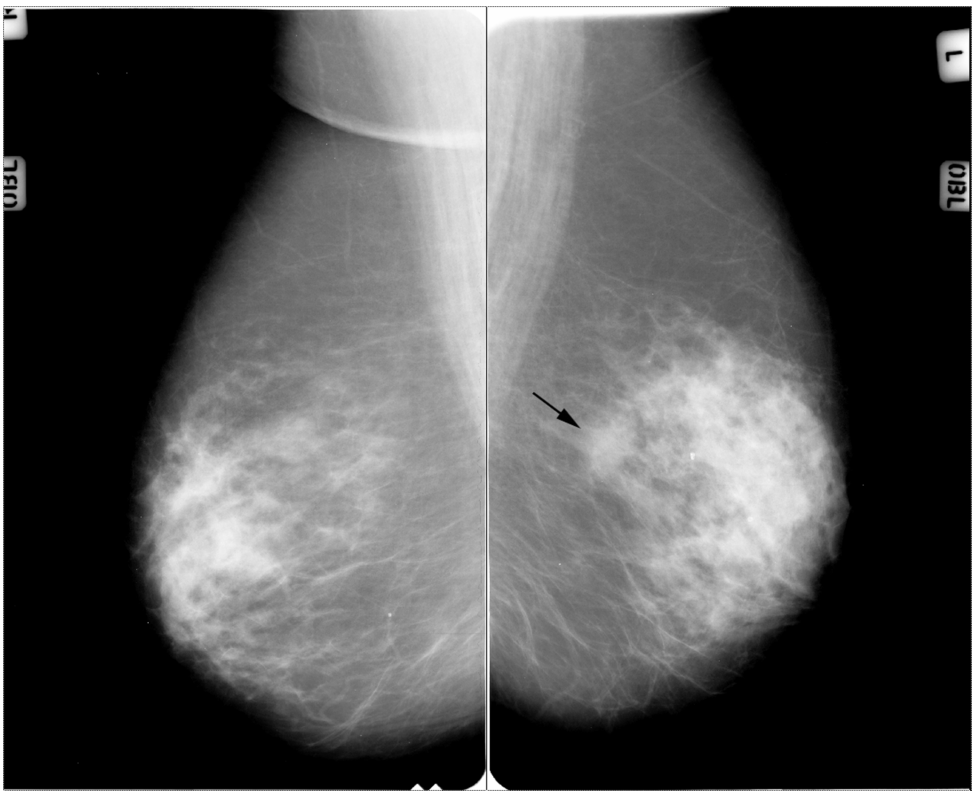


Figure 6.5: Mediolateral oblique mammographic views of a woman with an invasive ductal carcinoma indicated by the arrow. Seven of the nine readers correctly localized the cancer in both sessions, but rated their finding substantially more suspicious in the session with interactive CAD enabled, one reader only located the cancer correctly in the session where CAD was enabled, and one reader did assign a slightly lower rating to the cancer in the session with CAD.

Table 6.3: Mammogram reading times

	Average reading time per case (s)		
	Without CAD	With CAD	P value
Non-radiologists			
1	83.6 ± 47.0	111.5 ± 70.3	0.001
2	84.3 ± 59.2	67.7 ± 42.1	0.03
3	131.1 ± 65.1	129.5 ± 56.9	0.51
4	158.8 ± 68.1	146.0 ± 62.3	0.23
5	33.4 ± 29.6	35.2 ± 29.0	0.45
Average	97.0 ± 70.0	96.7 ± 67.4	0.97
Radiologists			
6	63.1 ± 45.6	58.9 ± 37.8	0.57
7	57.8 ± 31.7	70.8 ± 44.6	0.002
8	73.1 ± 44.1	73.1 ± 31.4	0.42
9	86.7 ± 52.1	88.6 ± 39.1	0.12
Average	70.0 ± 45.1	72.8 ± 39.8	0.02
Reader average	84.7 ± 61.5	85.9 ± 57.8	0.13

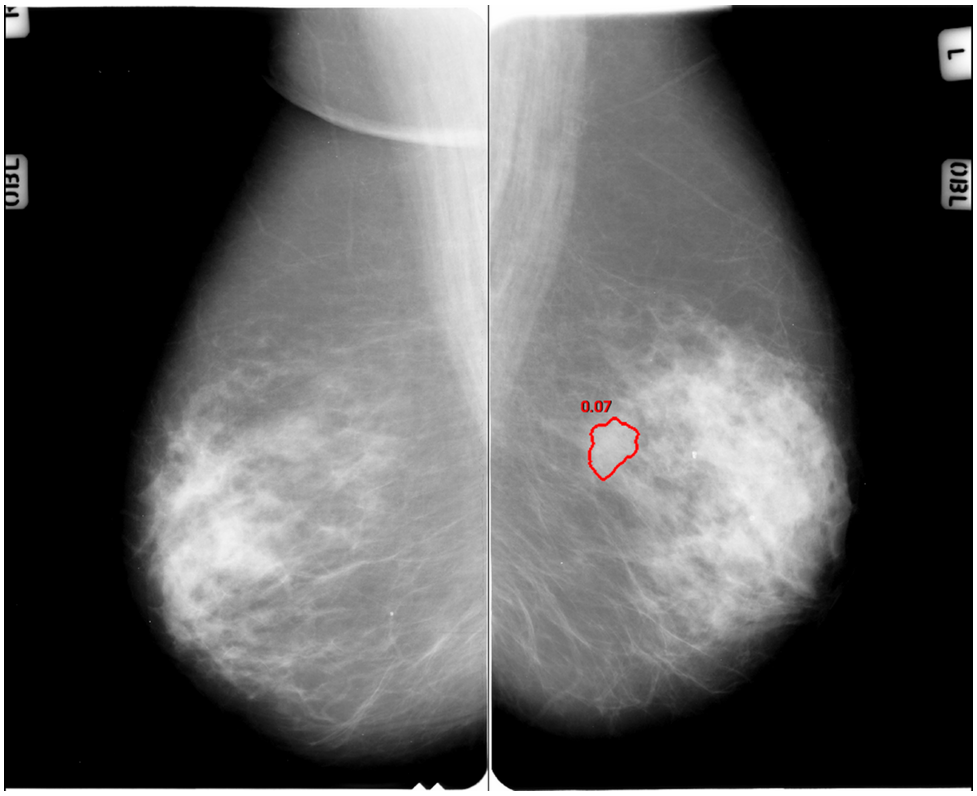


Figure 6.6: The same case as in Figure 6.5 with the activated CAD region. The *red contour* and a CAD score close to zero indicate a high probability that this is a cancer.

Table 6.4: Number of CAD regions queried[a]

	Queried CAD regions	Non-queried FP CAD regions	Non-queried, unreported TP CAD regions	Non-queried CAD regions regions but reported TP finding
Non-radiologists				
1	290	293	2	2
2	338	244	3	2
3	330	251	4	2
4	500	83	3	1
5	196	377	7	7
Average	330.8	249.6	3.8	2.8
Radiologists				
6	176	396	8	7
7	262	319	6	0
8	209	365	9	4
9	444	140	3	0
Average	272.75	305	6.5	2.75
Reader average	305	274.22	5	2.78

[a] There were 587 CAD regions in total; 546 false-positive CAD regions and 41 true-positive CAD regions

6.4 Discussion

Results of this study show that readers are able to improve detection performance when they use CAD for interpretation of mass lesions in an interactive way. The beneficial effect of CAD can be attributed fully to improvement of interpretation, because traditional CAD prompts to avoid perceptual oversights were not shown. The effectiveness was remarkable given that the readers in this study used the interactive system for the first time and had limited training. It is noted that in a previous experiment using a similar observer study design and data set no significant improvement with traditional CAD prompting was found when readers had limited training[199]. This suggest that for mass detection interactive CAD may be more effective than traditional CAD. This is in accordance with studies suggesting that interpretation errors are more common than perception errors[5,83]. Results obtained in this study show that readers are able to exploit the predictive power of CAD to improve their decisions. This may come as a surprise, because due to the large number of false-positives it is often believed that the performance of CAD for masses is much less than that of an experienced reader. It is noted, however, that in a previous study it was shown that the

performance of the CAD system was comparable to that of experienced readers when analysis was restricted to locations identified by the radiologists[85]. This is what counts in this study, because CAD results were only shown on regions probed by the readers. Interestingly, malignancy ratings of CAD were also used previously in the large CADET II trial[66] conducted in the UK, where the size of the CAD marks was used to represent the computed likelihood of cancer. Positive results of this trial could also be related to using CAD as decision support. The potential gain of using CAD for decision making was also demonstrated in a previous study, in which CAD information was independently combined with reader scores[83]. Results in this study confirm that by independent combination of reader scores with CAD performance can be improved (Table 6.2). On average, we found that the improvement in performance was larger when readers used CAD themselves than when CAD was independently combined with their scores. However, the difference was not significant. Interestingly, for one of the radiologists (number 8) detection performance decreased when using interactive CAD, whereas performance increased with independent combination. This may well be due to insufficient training. Readers need to learn how to weight CAD information in their decisions.

Table 6.3 shows the average reading times per reader for the sessions with and without CAD. We found that for the non-radiologists the average reading time was slightly reduced when they used CAD. For the radiologists the reading time increased less than three seconds on average with CAD. It seems that interactive use of CAD does not cost much extra time, because the information is presented at the moment the reader asks for it.

In the experiments we used a threshold to adjust the average number of CAD regions per image that could be activated. On average, there were two false-positives per normal image. In clinical practice the operating point of prompting systems for masses in mammography are often set to a level near 0.5 false-positives per image. We used more regions, because it was thought that in the interactive system more false-positives would be tolerable. Many of them are never activated, and if they are activated they are perceived very differently than traditional prompts. The radiologists queried far fewer false-positive CAD regions than the non-radiologists which may indicate they are more confident in their reading.

Interactive CAD is intended to aid the reader in decision making and will not help to avoid perceptual oversights. The success of the interactive approach may be explained by assuming that perceptual oversights do not occur frequently. In our study this appeared to be the case. On average only 5 (12.2%) of the true-positive CAD regions were not probed by the reader. Thus, in the reader study at most 12.2% of the cancers were overlooked, while none of them were reported in the original screening.

Results also show that on average 274.2 (50.2%) false-positive CAD regions were not activated, limiting the number of false-positives to which the readers are exposed. It is noted that the system can easily be extended by displaying the most suspicious, non-queried CAD regions as traditional prompts after the reading is completed.

In general, the response of the radiologists to the interactive CAD system was very positive and they preferred it to conventional CAD prompting systems. An advantage of the proposed system is that obvious false-positives of the CAD system are rarely shown, as the readers do not probe these regions. This may increase confidence in CAD.

In our study the reading conditions were less optimal than in screening practice, because a 4-megapixel color display was, instead of two 5-megapixel grayscale monitors commonly used in mammography. This might have a negative effect on the detection performance, especially for detecting microcalcifications. As microcalcification cases were not included in our study we do not believe that image quality influenced our study outcome. This is supported by a study from Kamitani et. al.[200] in which no significant differences were found between the observer performances for detecting breast cancer masses when performing soft-copy reading on 3-megapixel or 5-megapixel LCD monitors. Another limitation of our study is the absence of CC views in most cases. In the Dutch screening program, two-view mammography is not always performed at subsequent screens. Obviously, absence of additional CC views might affect the radiologists' detection performance. However, readers in our study are used to interpreting single view mammography. We would like to note that both limitations did not affect the difference in detection performance described in this paper, because the conditions were similar in the sessions with CAD and the sessions without CAD.

Participants in this study were not reading under normal screening conditions. It may be that their alertness, concentration and decision thresholds were affected by the knowledge that this study was a controlled laboratory experiment in which their decisions would be recorded and used in a study, and that the balance between cancer and normal cases was artificial. Because their assessments of the mammographic cases in this retrospective observer study would not affect patient care, their decisions could be different from that in an actual clinical setting. This effect has been described, among others, by Gur et.al.[201]. However, the reading conditions in the with-CAD and without-CAD were similar, and therefore the observed effect on detection performance can be attributed solely to the use of the interactive CAD system. Because we performed LROC analysis, decision thresholds did not affect study results.

As in many other studies, the sample was heavily weighted towards cancer cases. Not doing so would make this form of research extremely expensive. The effect on sensitivity and recall rates of radiologists using this interactive CAD system for real

life screening, can only be determined by a large randomized controlled trial in which radiologists use this system during routine use and for a substantial period[86]. Nevertheless, a laboratory study is generally a first step to demonstrate the usefulness of a CAD concept before a large trial is performed.

The readers participating in this study had different backgrounds and experience. We expect that when readers gain more experience with the system they will learn how optimize use of it. In addition, readers need to find out how to weight CAD information in their decisions, and we expect them to improve this when they gain more understanding of the strengths and weaknesses of the CAD software.

6.5 Conclusions

We found that in addition to using CAD in the traditional way to avoid perception errors, there is a large potential for using CAD as a decision aid to reduce interpretation failures. Results suggest that interactive CAD may be more effective than traditional CAD for improving mass detection without affecting reading time.

Interactive decision support versus prompting in mammography

7

Rianne Hupse, Maurice Samulski, Carla Boetes, Roel Mus, Ard den Heeten and Nico Karssemeijer

Original title: Computer aided detection of masses in mammography: interactive decision support versus prompting

Published in: Submitted to Radiology

Abstract

Purpose: To compare effectiveness of an interactive computer-aided detection (CAD) system, in which CAD marks remain hidden unless their location is queried by the reader and in which CAD marks are displayed with a suspiciousness score, to the effect of CAD prompts as used currently in clinical practice for the detection of malignant masses in mammograms.

Materials and methods: An observer study was conducted in which six certified screening radiologists and 3 residents read 200 cases (63 with a screen-detected malignant mass, 17 with a malignant mass missed in screening, 20 false positives from screening, and 100 normals) in two sessions. In each session, half of the cases were first read without CAD and subsequently with CAD prompts, while the other half of the cases were read with the interactive CAD system. Findings reported by the readers included location and a suspiciousness rating. For each reading mode, location receiver operating characteristic (LROC) curves were computed. Partial area under the LROC in an interval of low false-positive fractions (0 to 0.17, based on the observed false-positive rate) was used as a measure of reader performance. Differences in reader performance were analyzed using the Wilcoxon signed-rank test and the DBM-MRMC method.

Results: Reader sensitivity increased significantly ($p < 0.01$) when interactive CAD was used (58.5%) compared to both reading without CAD (51.2%) and reading with CAD prompts (51.1%). No significant difference was found in the number of unreported abnormal cases when mammograms were read with interactive CAD compared to reading with prompting CAD or to reading without CAD.

Conclusion: For detection of malignant masses in mammograms interactive use of CAD results as decision support may be more effective than the current use of CAD aimed at avoiding perceptual oversights.

7.1 Introduction

In breast cancer screening computer-aided detection (CAD) systems are used to avoid perceptual oversight of abnormalities in mammograms. The positive effect of CAD has been shown in several studies[66,69,87,188] but there are also studies in which no performance increase has been found[71,86,189]. For the detection of microcalcifications the performance of current CAD systems is relatively high, which is appreciated by most readers. However, there is less agreement about the benefit of using CAD for the detection of masses and architectural distortions. Many radiologists argue that CAD shows too many false-positive prompts to have a positive effect on mass detection[190,191]. On the other hand, it has been shown that masses are often missed due to incorrect interpretation[84,202] and that reader performance can be improved by retrospectively combining reader scores with the presence and probability of CAD mass markers[83,85]. These results motivated us to develop a CAD system aimed at aiding radiologists with interpretation of suspicious regions detected by themselves. In this interactive system, CAD marks are only displayed for queried regions and are accompanied by a suspiciousness score. In a recent study interactive use of CAD for detection of masses in mammograms was found to be effective[88]. In that study a commercial CAD system and a database of digitized film mammograms were used. The purpose of this study is to compare the effect of an improved interactive CAD system to the effect of traditional CAD prompts in a reader study. The study is carried out with full field digital mammograms (FFDM) randomly collected from a screening program.

7.2 Materials and methods

7.2.1 Study population

In this retrospective study all material was anonymized. Institutional review board approval was waived. All mammograms used in this study were acquired in a digital screening pilot project conducted in the period 2003-2008 at the Preventicon screening centre in Utrecht, the Netherlands[203].Women whose mammograms were included in the study completed a questionnaire at screening, in which they granted permission to use their mammograms for quality control and scientific and educational purposes. In the screening program, women in the age group 50-74 are invited to participate every two years. Processed digital mammograms were acquired with a Selenia FFDM system (Hologic, Danbury, CT). As common in the Netherlands, MLO and CC views are obtained at the initial screening, while on the subsequent only MLO views are made, unless there is an indication that obtaining the second view would be beneficial. All

mammograms were read independently by two radiologists, with referral based on consensus. For subsequent screenings, digitized prior film mammograms were available of the exam preceding the first digital screening exam.

7.2.2 Case selection

In the pilot project phase there were 1239 FFDM based referrals in which 202 cancers were detected with a mass or architectural distortion as the dominant sign of abnormality. All mammograms with abnormalities used in this study were annotated under supervision of an experienced reader who did not participate as reader in the observer study. These annotations were used as reference standard for validation of the reader scores. When a lesion was annotated, it was also assigned a subtlety score in the range 1 (obvious) - 5 (hardly visible). For the experiment 80 positive and 120 negative cases were selected from this series as described below. To make the study more representative for international standards we only selected cases in which both CC and MLO views were available. Cases in which the lesion was rated as obvious lesions (subtlety score of 1) and microcalcification cases were excluded. First we checked for digital screening mammograms acquired prior to detection in which the lesion was already visible. This yielded a total of 17 mammograms. From the remaining cases we randomly selected 63 mammograms from incident screening rounds with a detected cancer. For the negative cases, to make the set more challenging we included 20 false positive referrals. Cases were selected in which the radiologist reported a suspicious mass or architectural distortion but no malignancy was found, further examination did not include biopsies, and at least one negative follow-up screening mammogram was obtained. Obvious benign abnormalities were excluded. The remaining 100 negative mammograms were randomly selected from the non-referred digital mammograms in the pilot project and had at least one normal follow-up exam. We took care that the proportion of initial screenings was the same for positive (4 out of 80) and negative (6 out of 120) cases.

7.2.3 CAD and reading environment

The CAD system we developed for use in this study[204] was designed to detect malignant masses and architectural distortions and was trained on a large set of digitized film mammograms (11793 images containing 1853 malignant mass regions). Using a dedicated pre-processing module this system can be used for detection of abnormalities in FFDM cases[205]. The system detects suspicious locations in each mammogram and computes a contour for each location using an automatic segmentation proce-

dure[148]. For each detected region a suspiciousness score is computed based on features like spiculation, local contrast, size and shape. A special feature of the CAD system is that detected regions in the MLO and CC views of the same breast are automatically linked if they were classified as corresponding lesions based on similarity and location[51,79,204]. Like in other recent developed CAD algorithms[52,176,206] this linking information is used when computing the suspiciousness score for a region. In our CAD system the linking information is also used when the CAD results are displayed. Comparison with prior mammograms was not included in the CAD algorithm.

The reader study was performed using an in-house developed experimental reading environment for screening mammography[88], which includes hanging protocols for navigation between views and comparison of current and prior mammograms. Images were displayed using a 30-inch dicom calibrated color LCD panel (model FlexScan SX3031W; Eizo Nanao Technologies Inc., Hakui, Ishikawa, Japan) with a native resolution of 2,560 $\times$ 1,600. Actions of the readers are logged by the system to facilitate detailed analysis of the sessions.

CAD results could be viewed in two different modes: the traditional prompting mode and the interactive mode. In the prompting mode, once activated all CAD regions are shown by displaying their contours, without providing a suspiciousness score. To limit the number of false-positives, only prompts are shown for CAD regions with a suspiciousness score above a threshold. This threshold is computed on a separated data set in order to display on average 2 prompts in a normal case (4 images). This prompting mode with the threshold we used is similar to use of CAD in current clinical practice.

In the interactive mode, each single CAD mark remains hidden until activated. A reader can activate a CAD mark by clicking with the computer mouse on a mammographic region. If a CAD result is available at the queried location, the contour of this region is presented to the reader with its suspiciousness score. A CAD result is considered available if the queried location is inside the contour of the CAD region or if the distance between the queried location and the center point of the CAD region is less than 0.5 cm. Also view correspondence is used: if an activated CAD region is linked to a region in the other view, for this other region the contour and score are also shown.

To display a CAD result, the suspiciousness score computed for the region should be above a threshold. This threshold is chosen in such a way that on average 8 CAD regions are available in a normal case (4 images). More CAD results are made accessible in this way than in the prompting mode, but they are only shown when activated. The contour of the region is displayed in color using a continuous scale from yellow (less suspicious) to red (highly suspicious). A numeric value representing suspiciousness is also shown next to the contour, after converting the CAD output to a scale of 0

(not suspicious) to 100 (very suspicious). This conversion is done with a lookup table which matches the CAD output with reader scores obtained in a former experiment[88], in which a different dataset was used.

7.2.4 Observer study design

Nine readers, of which six were certified screening radiologists and three were residents, participated in the study. Before starting the study the readers were trained with a short training session to become familiarized with the interface of the system, and with the CAD result presentation. This was done by presenting a set of 20 training mammograms to be read with prompting CAD and a set of 20 training mammograms to be read with interactive CAD. The readers were informed that the study set did not contain microcalcification cases. They were also informed about the approximate proportion of the abnormal cases.

In the actual observer study, each reader read all cases in both modes in two sessions. In the first session 100 cases were read in the first mode (prompting CAD or interactive CAD) and subsequently the other 100 cases were read in the other mode. The second session was at least 4 weeks after completing the first session. The cases that were read with prompting CAD in the first session were read with interactive CAD in the second session, and the other way around. Five of the readers used prompting CAD first, the others used interactive CAD first.

To obtain sufficient data for analysis, radiologists were asked to report more findings than they would normally do in their screening practice. This means that they also had to score regions they would normally not refer. They were instructed that an average of about one finding per mammogram would be a good response. A finding was reported by moving a mass icon to a suspicious location. Each finding was numbered. If a finding was visible in both views, radiologists were asked to report the finding by moving two mass icons with the same index number to both locations. Readers assigned a suspiciousness score in the range 0-100 to each finding using a slider. For each case the readers also indicated if they would refer the case or not.

In the prompting CAD mode, prompts were initially not displayed. The radiologist was forced to score the case first without any help of CAD. Subsequently, CAD prompts were displayed and the radiologist was able to add or remove findings, change scores, or alter the referral decision. In this way, the effect of traditional prompting could be compared to reading without CAD.

7.2.5 Data analysis

From the collected observer data location receiver operating characteristic (LROC) curves were computed. Sensitivity was computed as the fraction of abnormal cases in which the reader had reported the mass or architectural distortion at the correct location in at least one of the views. The location of a finding was considered correct if its distance to the center of the reference standard was less than 2 cm. If a malignant lesion was reported with multiple findings, the finding with the highest score was used. For one case two malignant masses were present. The correct localized finding with the highest score was used for this case. The false-positive fraction was based on the findings with the highest reader score in each normal case. For each LROC curve, the mean true-positive fraction (MTPF) in a false-positive fraction interval ranging from 0 to 0.17 was computed. An interval containing low false-positive fractions was chosen because in screening the operating point at which radiologists work is at a low false-positive fraction. The choice of the interval was based on the fact that in this study 17% of the normal cases (20 of a total of 120) were false-positives in the original screening.

To compare the mean sensitivity in the defined false-positive fraction interval for different modes, significance tests were performed using the Wilcoxon signed-rank test. For this test the paired MTPF values for each reader were used. The Wilcoxon signed-rank test treats only readers as a random sample. At the moment there is no good significance test available to compare different values for MTPF in which both readers and data are treated as random samples. The JAFROC method[207] does only compute the area under the whole LROC curve, while in screening only the sensitivity at low false-positive fractions matters. The DBM MRMC method[208] treats both readers and data as random samples. Although this method does not take into account the location of findings we also performed this test.

We investigated the influence of CAD on the number of cancers that were not reported. This was done by counting the cases for which the reader did not mark the abnormality in any of the views. To compare this number for the different modes, significance tests were performed using the Wilcoxon signed-rank test.

In the interactive mode the average number of clicks with and without CAD response was computed. Clicks were only counted in normal cases, excluding the referrals, because in screening most cases are normal. Reading times were analyzed by computing the median reading time for the normal cases (not the cases false-positively referred in screening) for each reader. Median reading times were computed instead of average reading times because the average value was affected by some excessively long reading times caused by interruptions during the session.

7.3 Results

With the prompting mode threshold the sensitivity of CAD was 84% (67 of 80 cases detected), with on average 3.2 false-positive findings per normal case. When the interactive mode threshold was used the sensitivity was 91% (73 of 80 cases) with on average 8.2 false-positive findings per normal case. In the prompting mode, true-positive prompts were displayed on average in 10.7 of the abnormal cases that were not reported when reading without CAD. In the interactive mode, true-positive CAD regions were available for on average 12.3 cases that were not detected by the reader without CAD, slightly more because of the lower threshold. On average the readers re-

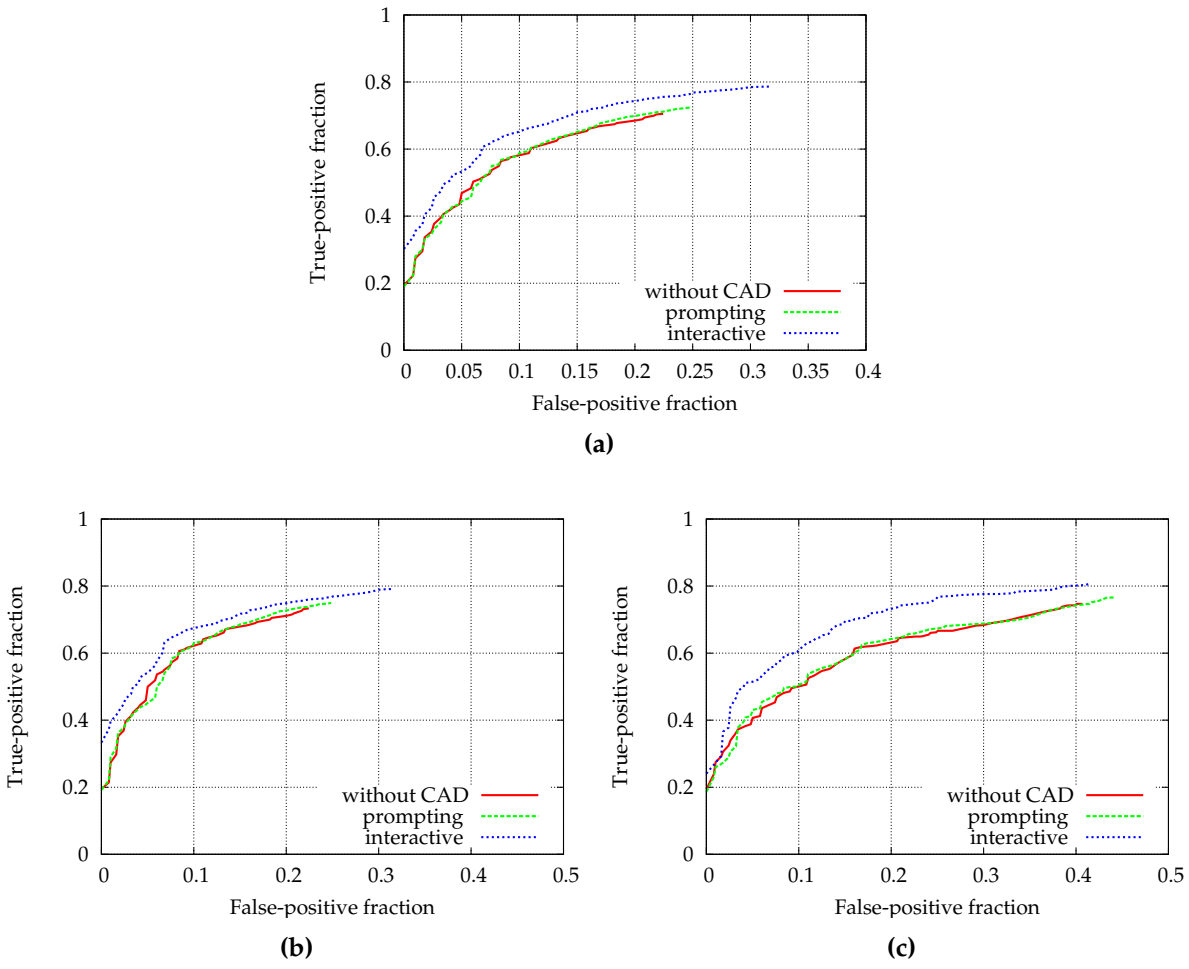


Figure 7.1: Location receiver operating characteristic (LROC) curves for reading without CAD, with prompting CAD and with interactive CAD. The curves are averaged over all 9 readers (a), the 6 certified screening radiologists (b) and the 3 residents (c).

ferred 17% of the normal cases in the mode without CAD. Figure 7.1a shows the LROC curves for the three modes (without CAD, prompting CAD and interactive CAD) averaged over the readers. For the whole interval of false-positive fractions, the sensitivity

obtained in the interactive CAD mode is higher than in the other modes. The LROC curves based on the results for respectively the 6 certified mammographers and the 3 residents are depicted in figure 7.1b and 7.1c. For most false-positive fractions, the effect of interactive CAD is higher for the residents than for the certified radiologists. The mean true positive fraction (MTPF) in the false-positive fraction interval from 0 to 0.17 is listed for each reader in table 7.1. For the mode without CAD the MTPF is given, for the other modes the increase in MTPF is given compared to the mode without CAD. For almost all readers the use of interactive CAD yielded a higher MTPF than the mode without CAD. On average the mean sensitivity increased from 0.512 (without CAD) to 0.585 (interactive CAD). This difference was significant ($p = 0.009$ for the Wilcoxon signed-rank test and $p = 0.003$ for the DBM MRMC method). When comparing interactive CAD to prompting CAD (average MTPF of 0.511), the increase was also significant ($p = 0.002$ for the Wilcoxon signed-rank test and $p = 0.003$ for the DBM MRMC method). For one of the certified screening radiologists the MTPF was slightly lower when interactive CAD was used compared to reading without CAD. This was the reader that achieved the highest MTPF in the mode without CAD.

All abnormalities were correctly localized by at least one reader in the mode without CAD. The number of unreported abnormal cases is listed in table 7.2 for each reader and each mode. For most readers (6 out of 9) the number unreported abnormal cases decreased when prompting CAD was used, for the other 3 no difference was found. On average the difference in unreported abnormal cases was 1.33 (significant, $p = 0.019$) for the prompting mode compared to reading without CAD. For the interactive mode the number of unreported abnormal cases increased for 4 radiologists (ranging from 1 to 2 cases) and decreased for 4 radiologists (ranging from 2 to 12 cases) compared to reading without CAD. Overall, disregarding the observer ratings, no significant difference in the number of unreported abnormal cases was found when reading with interactive CAD was compared to reading with prompting CAD or to reading without CAD.

The average number of clicks, in the interactive sessions, per normal case with and without CAD response is listed in table 7.3. There is a large variance in the number of times the readers queried CAD, which ranged between 0.3 and 13.1 per case. For each reader at least half of the clicks did not activate any CAD region. The median reading time per case for each reader is given in table 7.4. On average, the reading time increased by approximately 10 seconds when using prompting CAD or interactive CAD.

Table 7.1: Mean true-positive fraction (MTPF) in the false-positive fraction interval 0-0.17. For the mode without CAD the MTPF is given, for the other modes the increase in MTPF is given compared to the mode without CAD

	Without CAD	Prompting vs. without CAD	Interactive vs. without CAD
Mammographers			
reader 1	0.491	+0.006	+0.046
reader 2	0.515	+0.017	+0.117
reader 3	0.677	−0.026	−0.018
reader 4	0.539	−0.017	+0.036
reader 5	0.503	−0.006	+0.071
reader 6	0.520	+0.012	+0.117
average	0.541	−0.002	+0.062
Residents			
reader 7	0.440	+0.009	+0.169
reader 8	0.399	−0.001	+0.117
reader 9	0.528	+0.001	+0.004
average	0.456	+0.003	+0.097
reader average	0.512	−0.001	+0.073

Table 7.2: Number of unreported abnormal cases. In parentheses the increase is given compared to the mode without CAD

	Without CAD	Prompting	Interactive
Mammographers			
reader 1	13	11(-2)	14(+ 1)
reader 2	9	8(-1)	11(+ 2)
reader 3	10	10(0)	11(+ 1)
reader 4	12	11(-1)	13(+ 1)
reader 5	22	22(0)	22(0)
reader 6	24	23(-1)	12(-12)
average	15.0	14.2	13.8
Residents			
reader 7	23	17(-6)	14(- 9)
reader 8	17	17(0)	14(- 3)
reader 9	13	12(-1)	11(- 2)
average	17.7	15.3	13.0
reader average	15.9	14.6	13.6

7.4 Discussion

We found that reader performance for the detection of malignant masses and architectural distortions increased when CAD results were interactively displayed compared to regular prompting or reading without CAD. Readers also preferred the interactive system. This can be explained as follows. In the interactive mode marks remain hidden unless corresponding regions are probed. As most radiologists only probe a limited number of regions, and only those they are interested in, less false-positives are displayed. Because display is initiated by the reader CAD does not disrupt the reading process. CAD suspiciousness scores are used to aid with interpretation and may change the initial opinion of the reader. As the interpretation of the reader and the CAD ratings are correlated the readers gain more confidence in CAD than with the use of prompts.

More CAD marks were available in interactive reading mode. Therefore, findings with a relatively low suspiciousness score could be activated that were not displayed in the prompting mode. In this way, the readers could use CAD as interpretation support for subtle lesions for which they were not sure whether to recall or not. Generally, regions marked by CAD with low suspiciousness can support readers in their decisions

Table 7.3: Number of clicks per normal case with and without CAD response

	With CAD response	Without CAD response	Total
Mammographers			
reader 1	1.7	3.0	4.7
reader 2	2.2	3.8	6.0
reader 3	4.0	9.1	13.1
reader 4	1.1	2.1	3.2
reader 5	1.1	1.3	2.4
reader 6	0.2	0.2	0.3
average	1.7	3.3	5.0
Residents			
reader 7	0.6	0.9	1.5
reader 8	0.2	0.4	0.7
reader 9	2.1	6.1	8.2
average	1.0	2.5	3.5
reader average	1.5	3.0	4.5

not to recall. One of the advantages of the interactive system is that such CAD marks are not experienced as false-positives by the readers in the interactive system, but they rather strengthen the confidence of the readers in the system.

One might argue that a disadvantage of interactive CAD is that perception oversight errors are not avoided. We did not use eye-tracking methods in our study and therefore we do not know which regions were inspected. However, no significant difference was found in number of unreported cancer cases when interactive CAD was compared to prompting CAD. This suggests that oversight was not a major cause of missing cancers in our study. The number of unreported cancer cases even decreased for some readers in the interactive mode. This might be explained by the lower threshold we used, due to which more CAD marks were available in interactive mode CAD, which might have encouraged some radiologists to report more subtle abnormalities. We also found that for three out of the four radiologists who reported less cancer cases in the interactive mode, the mean sensitivity at low false-positive fractions increased. This suggests that the influence of interactive CAD on the given reader scores had more effect on reader performance than the increase of unreported cancers.

In our study we used a 4-megapixel color display. Although this display has less spacial and gray value resolution than displays used in clinical practice, it is not ex-

Table 7.4: Median reading time (sec./case). In parentheses the increase is given compared to the mode without CAD

	Without CAD	Prompting	Interactive
Mammographers			
reader 1	29	40(+11)	27(- 2)
reader 2	56	68(+12)	75(+19)
reader 3	42	58(+16)	61(+19)
reader 4	52	63(+11)	67(+15)
reader 5	34	44(+10)	57(+23)
reader 6	30	35(+ 5)	34(+ 4)
average	41	51(+11)	54(+13)
Residents			
reader 7	34	44(+10)	27(- 7)
reader 8	42	51(+ 9)	70(+28)
reader 9	51	59(+ 8)	41(-10)
average	42	51(+ 9)	46(+ 4)
reader average	41	51(+10)	51(+10)

pected that this will have influenced the detection of masses. A study from Kamitani et al.[200] showed no significant differences in observer performance for detection of masses between the use of a 3- or a 5-megapixel monitor. We asked the readers their opinion about the quality of the mammogram display and they responded that they found that the quality was excellent for mass detection.

Regarding reading times, we found that the use of CAD (interactive or prompting) lengthened the reading time by approximately 10 seconds. Due to the sequential scoring of each case without CAD and with prompting CAD, the reading time for prompting CAD could only increase compared to the mode without CAD. In the interactive mode, the variance in number of CAD queries was large between readers. Some readers reported they spend more time exploring the CAD results in the interactive system out of curiosity, which will have increased the reading time. Therefore, we expect that over time reading time will be reduced. It is noted that in an earlier study we found no increase in reading times with interactive CAD[88].

In a retrospective study readers might perform differently compared to normal screening[201]. Reasons might be that decisions do not affect patient care, there is a competition element to perform better than colleagues, and readers know the dataset contains more cancers than in screening practice. We compared results of this study

to the performance obtained in the original screening for the set of cases we selected. In the original screening, 79% of the abnormal cases in our dataset were detected (63 out of 80) at a false-positive rate of 17% (20 out of 120). Our results show an average sensitivity of 69% at a false-positive rate of 0.17 for certified radiologists reading without CAD. These are single reader results while in original screening each case was read by two radiologists with differences of opinion resolved in consensus. According to a meta-analysis[86], the cancer detection rate increases by approximately 10% when mammograms are double read. Taking this into account, we expect that when double reading would have been used in this study similar performance would have been obtained as in the original screening. This indicates that the readers in the observer study were not behaving different than in practice. Further, on average 17% of the negative cases were referred by the readers in this study. This is similar to the percentage of cases in our study set that were false-positively referred in screening.

In this study the effect of prompting CAD may be underestimated. Our results differ from results obtained in prospective studies in which prompting CAD had a positive effect on reader performance[66,69,87,188]. It might be that in a retrospective observer study like ours less search errors are made and that therefore prompting CAD has less effect. Nevertheless, in that case the fact that no difference was found between reading without CAD and with prompting would be the result of a higher performance for reading without CAD and not of a lower performance for reading with prompting. Another reason might be that in the prompting mode the readers scored each case before and after the CAD prompts were displayed. Readers might be less inclined to act on CAD prompts when a decision was already made without CAD.

It is a limitation of the study is that the size of the effect of interactive CAD we found cannot be translated easily to screening practice. We selected a challenging set of cases for this study, in which the proportions of normal and abnormal cases were different than in screening practice. Microcalcification cases in which no mass or architectural distortion was visible were excluded and normal cases were enriched with difficult cases referred in the original screening. Reader performance is very dependent on the subtlety of the cases in the study set. However, we used the same study set for each mode and therefore we believe that the relative differences between reading without CAD, with prompting, and with interactive CAD are valid.

In conclusion, for detection of malignant masses in mammograms the interactive use of CAD results as decision support may be more effective than the current use of CAD aimed at avoiding perceptual oversights.

Computer-aided detection as a decision aid in chest radiography

8

Maurice Samulski, Peter Snoeren, Cornelia Schaefer-Prokop, Bram van Ginneken and Nico Karssemeijer

Original title: A Novel Method for Presentation of Computer-aided Detection Results in Chest Radiographs

Published in: To be submitted

Abstract

A method for presenting computer-aided detection results is proposed in which readers probe image locations for decision support. This could be an alternative to the current method of displaying CAD prompts. While not aimed at avoiding visual search errors, the new approach has the advantage that false positives of CAD are not distracting the reader and that decision errors may be reduced. The aim of the study is to investigate interactive CAD presentation in a lung nodule detection task, and to compare its effect on readers to the effect of prompts. We used 223 chest radiographs from the public JSRT database, including 130 cases with lung nodules. Six readers participated in an observer study in which cases were interpreted unaided, with prompting CAD, and with interactive CAD. Readers reported locations of findings and rated these on a continuous scale. For analysis localization receiver operating characteristic (LROC) was used. Mean sensitivity was computed in an interval of false-positive fractions less than 10%. With CAD prompting, mean sensitivity of the readers increased significantly from 35.2% to 42.8%. When using interactive CAD, the performance of the average reader increased significantly to 49.5%. Using CAD interactively as a decision aid can improve readers' detection performance significantly compared to the traditional use of CAD prompts, in particular at low false positive rates.

8.1 Background

It is generally believed that perception errors in radiology may be reduced if a computer aided detection system displays prompts on potential abnormalities it has detected[209]. In this way it can be avoided that these abnormalities are overlooked. When CAD is very sensitive, prompts may also reduce reading times, as radiologists can then use CAD as a reliable guide to quickly find relevant image regions that need careful inspection. Many studies have shown that CAD prompts can have a positive effect on detection performance[86]. However, in practice CAD technology is not yet widely used. Radiologists are generally not convinced that the technology is effective at its current stage. The major complaint is that CAD produces too many false positives. Only in mammography CAD systems are used on a large scale[210].

In this study we investigate a new concept for presenting CAD results, which aims at making CAD more tolerable and more effective. The study is motivated by the fact that in radiological detection tasks the reader does not only have to find potential lesion locations, but also has to decide whether abnormalities are true lesions or not and if they are actionable. CAD prompts only help with the search task and leave the second task to the reader. This is remarkable, because there is no evidence that the search task is more difficult for readers than the interpretation task. In fact, some experimental studies suggest that the reverse is true in common radiological tasks. In addition, it seems that computers are better in lesion interpretation than in searching for them. Lesion characterization studies have been reported in which the computer performs equal or better than experienced radiologists[211–214], while in standalone lesion localization tasks CAD systems generally perform much worse than human observers, due to a higher false positive rate. Therefore, we investigate how CAD can be used to improve the decision stage, when the potential abnormality already has been localized.

CAD algorithms developed for detection of abnormalities do not need to be changed to use them as a decision aid. CAD results should only be presented in a different way to the readers. We propose an interactive method, in which CAD results associated with a specific image location are only shown if the reader queries the location. The idea is that if the reader is in doubt whether an actionable abnormality is present at the queried location CAD may help to make the right decision. CAD results displayed to the reader may depend on the application, but an essential component should be a score representing the probability that a true lesion is present. Such a score is commonly computed by CAD systems. In prompting systems this score is used to determine if a prompt should be shown or not. Only when the score exceeds a predefined threshold a prompt is shown. In this study, a contour representing the outline of the lesion detected by CAD is shown in addition to the suspiciousness score. CAD results

are displayed only when the reader probes a location that holds CAD information with a score exceeding a threshold.

To study the proposed CAD presentation method we focus on lung nodule detection in chest radiography. A preliminary report on this work was published previously[204]. Lung cancer is a leading cause of death worldwide, and accounts for 1.4 million deaths in 2008[215]. Chest radiography is the most common imaging technique for the diagnosis of pulmonary diseases, mainly due to low cost and short examination time[94]. However, it has been shown that the detection of pulmonary nodules at an early stage in chest radiographs is an extremely difficult task for radiologists and there are many papers reporting about radiological error in chest radiography[95–99]. Of particular interest here are studies from the perception literature. In a well known experiment with briefly flashed chest radiographs (a few tenths of a second), Kundel and Nodine[10] showed that visual search for pulmonary nodules begins with an almost immediate and global response with surprisingly high performance. More recently, it has been shown that the majority of errors in the detection of pulmonary nodules is related to recognition and interpretation, whereas only a minor fraction can be explained by incomplete search patterns[4,5,216].

Lung nodule detection in chest x-rays was one of the first applications for which CAD was developed. Commercial systems providing CAD prompting exist but are not yet widely used. Several research groups have been investigating the performance of CAD systems for the detection of lung nodules on chest radiographs[103–105], and the effect on the radiologists' performance[107,108,119,217]. Current clinical CAD systems for the detection of lung nodules in chest radiographs are based on the idea that prompts will help avoid perception errors. Results from perception studies reported above suggest that providing CAD prompts may not be the most effective approach to reduce nodule detection errors and that the interactive presentation method may lead to better results. In a previous study we already demonstrated that interactive display of CAD had a positive effect on reader performance in mammography[88]. In this study the purpose is to make a direct comparison between the effect of prompts and interactive use of CAD.

8.2 Method

8.2.1 Data set

All chest radiographs used in this study were selected from the publicly available JSRT database of the Scientific committee of the Japanese Society of Radiological Technology[218]. It consists of 247 posteroanterior chest radiographs; 154 images are abnormal containing a solitary pulmonary nodule, and 93 are normal. The images were digitized

from original screen-film images with a $0.175\,mm$ pixel size, $2,048 \times 2,048$ matrix size, and a gray scale depth of 12 bits. The cases were collected from 13 medical centers in Japan and one in the United States. The radii of the nodules range from $2.5\,mm$ to almost $30\,mm$, and the median value is $7.4\,mm$. The nodules were divided into five subtlety categories based on the consensus of three chest radiologists: extremely subtle ($n = 25$), very subtle ($n = 29$), subtle ($n = 50$), relatively obvious ($n = 38$), and obvious ($n = 12$). More than two-third of the cases were considered subtle. By using a public database for our study comparison of results to other studies is facilitated.

8.2.2 CAD system

The CAD results of a commercially available CAD system OnGuard™5.0 (Riverain Medical®, Miamisburg, Ohio, USA) were used. For the JSRT database, this CAD system prompted 386 locations: 105 on normal images and 281 on abnormal images. 43 images have no CAD prompts at all: 31 normal images and 12 abnormal images. For each CAD prompt Riverain Medical provided a computer estimated malignancy score between approximately -1.0 and 9.0, where a higher number indicates a higher likelihood of malignancy.

This internal CAD score is an abstract measure for suspiciousness that is hardly, or not, to be understood by humans. Therefore, we converted this abstract score into an interpretable measure, namely the probability that the prompted CAD location is inside a truth region, i.e., the probability that it is a true-positive (TP). To prevent bias, the computations were done by a leaving-one-image-out method. The converted CAD score could both be displayed as a number $\in [0, 1]$ and as color coding. See the appendix for a detailed description of the conversion procedure.

8.2.3 Workstation

For the purpose of this study, a previously developed workstation is used that has basic functionality such as zooming, image manipulation, local contrast enhancement and grayscale inversion tools. The brightness and contrast settings were easily adjustable and were set in advance for optimal efficiency. The chest radiographs were viewed on a 30 inch color LCD panel (model FlexScan SX3031W; Eizo Nanao Technologies Inc., Hakui, Ishikawa, Japan) with a native resolution of $2,560 \times 1,600$, a contrast ratio of $900 : 1$, and an intensity of $260\,cd/m^2$. On the workstation (see Figure 8.1) the presence of CAD prompts can be queried interactively by clicking on suspect regions in the chest radiograph using the computer mouse. When a location in the chest radiograph is queried, the workstation checks if a CAD mark exists on that location. If a CAD

mark is available, a circle is displayed with the computer-estimated probability. The circle is colored based on the probability of cancer and ranged continuously from red to green, for respectively high to low probabilities.

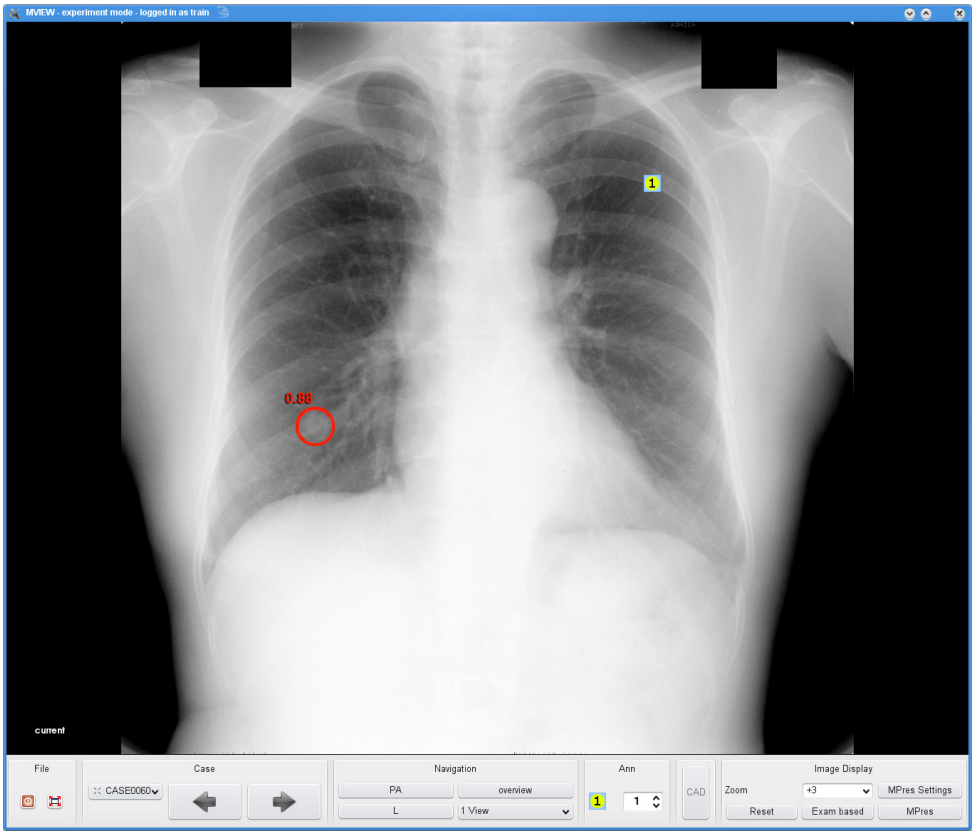


Figure 8.1: Snapshot of the CAD workstation used in the observer experiment. A snapshot made during a session with *interactive CAD prompts*. The label with "1" on the left lung is an annotation made by the reader. The (red) circular region on the right lung is a CAD prompt given after that the subject queried that location with a pointer device. The display provides a CAD computed probability of 0.88 that this region is a TP, while the probability also determines the color of the prompt. After all regions a user wants to report are marked, readers have to provide ratings of the findings with a slider before they can continue with the next case.

8.2.4 Experimental Design

In this study two conditions were investigated, traditional prompting CAD in a sequential design and interactive CAD in a concurrent design. In the traditional prompting CAD session, the readers were first asked to mark suspect regions without access to CAD. Then the marked regions were scored using a continuous rating scale between 0 (not suspect) and 100 (highly suspect). After the reader marked and rated the case, all available CAD prompts for the image were displayed without any information

about how suspicious the prompts are. The reader can accept the prompts as relevant and report them, or dismiss them. In addition to adding new findings, the reader has to confirm existing findings reported before the CAD prompts were displayed. The reader was also allowed to modify ratings of the annotations previously made without assistance of CAD.

In the interactive CAD session, areas of interest or suspicion can be sampled interactively by the reader. If a CAD prompt is available at a queried location then it is displayed by a color-coded circle (green: not suspect; red: highly suspect) and a number giving the probability that the prompt is a TP (see Figure 1). Suspect regions are marked by the reader while reading, and scored with ratings between 0 (not suspect) and 100 (highly suspect) after all regions are marked.

The experiment started with a training phase. For this purpose, a representative set of twenty-four chest radiographs were selected from the JSRT database containing nodules from all 5 subtlety categories. First, subjects were able to see the twenty-four images together with their ground truth and CAD regions. This phase was meant to demonstrate the subtlety of typical pulmonary nodules in the database and to familiarize non-experienced readers with the nodule detection task. Thereafter, the same twenty-four images were presented under the two experimental conditions to become acquainted with the user-interface and to establish a strategy for scoring the suspiciousness of annotations. The remaining 223 JSRT images were used for the actual experiment. Three random, mutually exclusive sets of images were constructed, consisting of about 75 images each. The sets were stratified by a subtlety rating that is available for all abnormal JSRT images. Each image was seen twice by a subject; once in each of the two experimental conditions. A total of six sessions were done, i.e., two conditions $\times$ three image sets. The order of image sets and reading methods were balanced over the subjects to minimize learning bias. We took care that no images were seen twice in the first three sessions, and only in the last three sessions images are presented for the second time (with the alternative reading method). To decrease a potentially negative effect of remembering cases we demanded a pause of at least a week between the first and the second three sessions.

The subjects' task was to find as many as possible abnormalities with as least as possible mistakes. The subjects were told to treat CAD as an auxiliary tool, a second reader, and not as the leading system. The subjects knew the prevalence of images with pulmonary nodules and they also knew the performance of the CAD system, which was summarized by a FROC curve and the fractions of lesions per probability rating.

8.2.5 Readers

Six readers participated in this experiment. All subjects were non-radiologists involved in radiological research projects. Two of them had extensive experience with reading chest radiographs, two had limited experience, and two had no experience with interpreting chest images. All subjects had normal, or corrected to normal vision and were familiar with the purpose of the experiment.

8.2.6 Reading times

There was no limitation on the reading time. During the reading sessions reading times per case were automatically recorded. When a subject did not move the mouse and did not do any other action on the workstation for more than 2 minutes this was recorded as idle time in the experiment data file. This idle time is subtracted from the reading time on the basis of the assumption that these excessively long idle times were the result of interruptions during the session. The average reading time per case and its standard deviation was computed for every reader for all three reading modes. Paired reading times were compared by Wilcoxon signed rank testing. A p value of less than 0.05 was considered to indicate a statistically significant difference.

8.2.7 Performance analysis

We used localization receiver operating characteristic (LROC) analysis for evaluation the readers' performance in the detection of lung nodules on chest radiographs. To determine a LROC, the decision threshold is varied and the correct localization fraction is plotted as a function of the fraction of normal cases that were recalled.

An marked location or CAD location is considered a true-positive (TP) when it is less than $2\,cm$ from the center of a pulmonary nodule, otherwise it is considered a false-positive (FP). In clinical routine only few chest radiographs have suspicious nodules. Therefore, the left part of the LROC curves represents the most relevant range. Sensitivity for higher false positive levels is undefined, as readers reported findings in only a fraction of the normal cases. The performance is computed as the mean correct localization fraction in the false-positive fraction interval ranging from 0 to 0.1. This interval was chosen because in screening programs radiologists usually have recall rates below 10%.

We performed receiver operating characteristics (ROC) analysis to compare the performance of the readers in this study to the results of twenty radiologists that were published in Shiriashi et. al.[218] for the same data set, without the aid of CAD. In addition, we evaluated readers performance using jackknife free-response receiver operat-

ing characteristics software (JAFROC 4.0, Dev P Chakraborty, 2011), which is used to estimate statistically significant differences between alternative FROC curves.

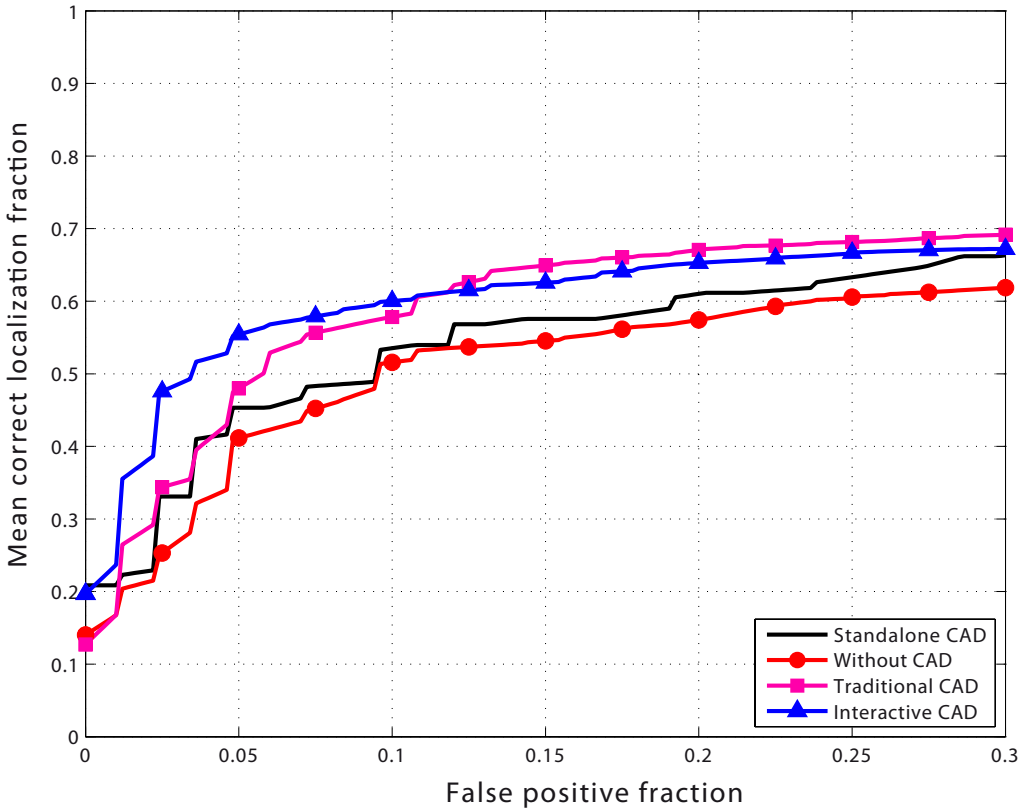
8.3 Results

The detection performance of the readers and the CAD system is given in Figure 8.2. In clinical routine only few chest radiographs have suspicious nodules. Therefore, the left part of the LROC curves represents the most relevant range. Sensitivity for higher false positive levels is undefined, as readers reported findings in only a fraction of the normal cases. As in previous research[88], the performance is computed as the mean correct localization fraction in the false-positive fraction interval ranging from 0 to 0.1. With traditional CAD, the performance of the average reader increased at a low false-positive range from 35.2% to 42.8%. When using interactive CAD the performance of the average reader increased from 35.2% to 49.5% in the same false-positive range. The differences between the columns in the table are significantly different from zero (sign test: $p < .05$). The performance of the CAD system, 39.1% in the false-positive fraction interval 0 to 0.1, appeared to be surprisingly good. CAD stand alone was better than the average reader without CAD support.

Table 8.1: Chest radiograph reading times. Average reading times per case (seconds). Reading times are displayed as mean ± standard deviation.

Reader	Without CAD	Traditional CAD	Interactive CAD
r1	16.7 ± 7.7	28.5 ± 13.6	16.1 ± 6.7
r2	48.2 ± 34.9	61.9 ± 42.9	39.2 ± 24.9
r3	48.6 ± 35.0	57.9 ± 38.3	40.1 ± 26.8
r4	33.6 ± 24.1	47.2 ± 30.3	41.9 ± 25.9
r5	31.9 ± 27.4	42.7 ± 32.5	54.9 ± 38.2
r6	25.3 ± 19.5	33.4 ± 23.4	26.7 ± 16.9
average	34.0 ± 28.8	45.2 ± 33.8	36.5 ± 28.0

The figure-of-merit values obtained with JAFROC for the average reader shows that the performance of the readers is 0.673 without using CAD. Using traditional CAD, the performance increased to 0.713 ($p < 0.001$). Using interactive CAD, the performance increased to 0.725, which was statistically significant different from using no CAD ($p < 0.001$). However, there was no statistically significant difference between both CAD modes. This can be explained by the fact that JAFROC evaluates the area under the whole AFROC curve. It was shown with LROC analysis that using interactive CAD had a more beneficial effect on the detection performance within the clinically relevant



(a) Average LROC curves

Reader	Without CAD	Traditional CAD	Interactive CAD
r1	0.286	0.401	0.466
r2	0.394	0.485	0.555
r3	0.376	0.405	0.477
r4	0.427	0.479	0.485
r5	0.302	0.444	0.520
r6	0.324	0.355	0.468
average	0.352	0.428	0.495

(b) Individual Performances

Figure 8.2: LROC curves. Four LROC curves are given in Figure 8.2a. Except for the stand alone CAD performance, the curves are averages over all subjects. Table 8.2b gives the individual performances of the subjects. These are computed as the mean correct localization fraction in the false positive fraction ranging from 0 to 0.1. Notably, the stand alone detection performance of the CAD system is 0.391.

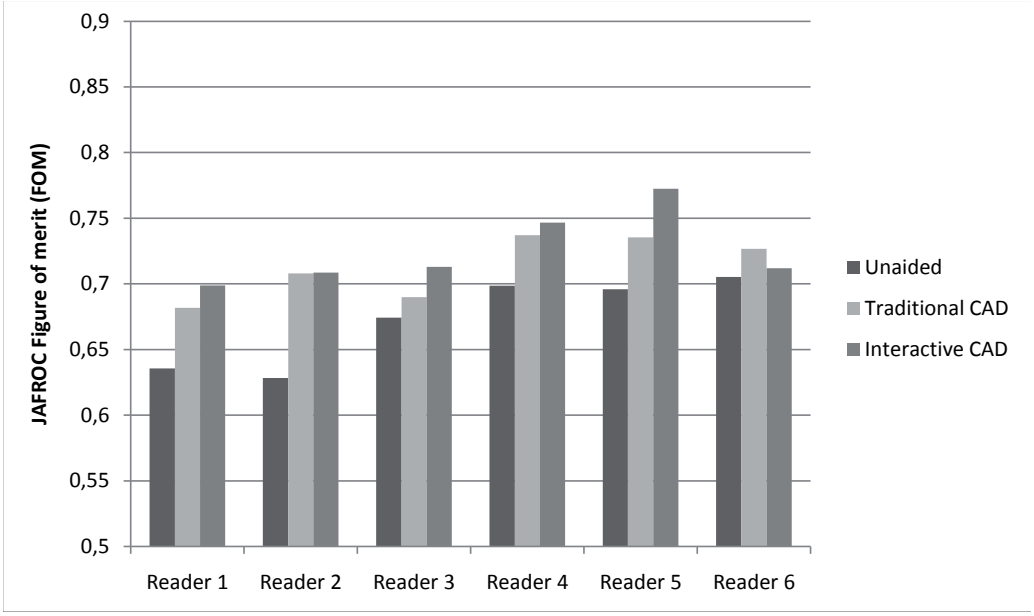


Figure 8.3: Jackknife receiver operating characteristic curves.

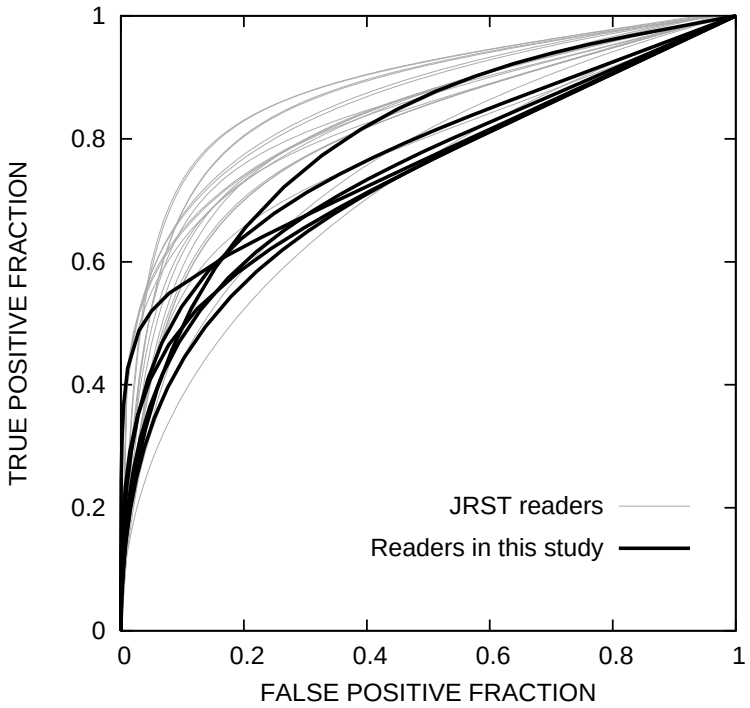


Figure 8.4: Receiver operating characteristic curves for all observers. The thin gray lines represent the radiologists from the JSRT study[218], the thick black lines represent the readers in the current study. Note the large variation in detection performance among readers.

false-positive fraction interval than using traditional CAD. Using traditional CAD led to a higher sensitivity, but not at a satisfactory recall rate.

In Figure 8.4 the receiver operating characteristic curves are plotted for the twenty radiologists from the study from Shiriashi et. al.[218], and the ROC curves for the readers that participated in this study.

The average time to read a case without CAD was $34.0 \pm 28.8s$. The average reading time increased to $45.2 \pm 33.8s$ when the traditional CAD prompts were activated and the reader re-evaluated his findings. In the interactive CAD session, the average reading time was $36.5 \pm 28.0s$ (Table 8.1).

8.4 Discussion

Our results indicate that interactive use of CAD can benefit readers in interpreting lung nodules on chest radiographs. The performance of the readers with interactive CAD is significantly better than traditional prompting CAD in a clinically important recall interval below ten percent. This confirms the hypothesis that readers take more advantage of CAD as interpretation aid than detection aid.

In the interactive mode, less false-positives are exposed to the readers as only a limited number of regions are queried. Because the regions are displayed on reader's request while reading the case, the reading process is not disrupted. Moreover, the reading time with interactive CAD is significantly smaller ($p < 0.001$) than reading with traditional CAD. On average, it took only 2.5 seconds longer to read a case using interactive CAD compared to unaided reading of a case.

The experience level of the readers could influence how beneficial CAD is: less experience will lead to a potentially bigger increase in detection performance. The performance of the readers that participated in this study is comparable to that of the lower range of general radiologists, as was shown in Figure 8.4, and a large variation in the detection performance of the observers can be observed. On average, the detection performance of the readers was lower than the radiologists that participated in the JSRT study[218]. We are planning to conduct an observer study with experienced chest radiologists to evaluate the effect of interactive CAD on their detection performance.

Further improvements might be achievable by adding traditional CAD prompts in the interactive mode, especially if they are rated highly suspicious by CAD and are on regions that were not interactively inspected by the reader, to help overcome occasional perception errors. Especially when lesions are obscured by other tissue, e.g., by the heart, interactive CAD can be inadequate, because lesions are harder to find in those regions and thus often not queried. In those cases a hybrid presentation of CAD

results may be helpful. In such a system, obscured CAD findings could be presented traditionally and the unobscured CAD findings could be queried interactively. An other possibility would be offering cases that were not recalled with very suspicious non-inspected CAD results at the end of the reading session.

Our study was limited in that the participants in this study were not reading under normal screening conditions. This was a controlled laboratory experiment, in which the participants knew the balance between cancer and normal cases and that their decisions would be monitored. They were also explicitly asked to search for lung nodules, whereas in clinical practice chest radiographs are also often requested for other reasons than lung cancer screening. Because their assessments in this observer study would not affect patient care, their decisions could be different from those in an actual clinical setting[201]. The reading conditions in the sessions with sequential traditional CAD and interactive CAD were similar, and therefore the observed improvement of detection performance can be attributed solely to the use of interactive CAD.

Acknowledgments

The authors would like to thank the readers that participated in the observer study, and Riverain Medical for providing the CAD results.

Appendix

In this section we explain in detail how we have computed the probabilities that were shown to the reader for each CAD prompt. Suppose we have a set of images with given truth regions. Furthermore, for each image we have a number of CAD regions; these are locations plus some corresponding raw scores. A CAD score is in general an abstract measure for suspiciousness that is hardly, or not, to be understood by humans. Therefore, we converted this abstract score into a interpretable measure, namely the probability that the prompted CAD location is inside a truth region, i.e., a *TP*.

The probability for a TP or a FP can be modeled by

$$\Pr\{r|s,\alpha,\beta\} = \begin{cases} \Psi(\beta(s-\alpha)) & \text{for } r = \text{TP} \\ 1- \Psi(\beta(s-\alpha)) & \text{for } r = \text{FP} \end{cases}$$

where r is the class of the CAD region, s the raw CAD score, and Ψ is a sigmoid function, e.g., a logistic function

$$\Psi(t) = \frac{1}{1+\exp(-t)},$$

(used in this paper) or an error function. If α and β are somehow computed, say they have fixed values α^* and β^*, then

$$\Pr\{\text{TP}|s\} = \Pr\{r = \text{TP}|s, \alpha^*, \beta^*\} = \Psi\left(\beta^*\left(s - \alpha^*\right)\right).$$

The problem is now to find the parameters α^* and β^*.

Assuming that all regions are independent (and that is not a bad assumption, as images are independent), we can compute the following probability ($\mathbf{r}$ is a set of true-positive and false-positive CAD regions, and $\mathbf{s}$ are the corresponding raw CAD scores)

$$\begin{aligned}\Pr\{\alpha, \beta|\mathbf{s}, \mathbf{r}\} &= \frac{\Pr\{\alpha, \beta|\mathbf{s}\}\Pr\{\mathbf{r}|\mathbf{s}, \alpha, \beta\}}{\Pr\{\mathbf{r}|\mathbf{s}\}} \\ &= \frac{\Pr\{\alpha, \beta|\mathbf{s}\}\prod_i \Pr\{r_i|s_i, \alpha, \beta\}}{\Pr\{\mathbf{r}|\mathbf{s}\}}\end{aligned}$$

$\Pr\{\alpha, \beta|\mathbf{s}\}$ is the prior pdf. We could have defined some function for this, but we leave it as a constant. $\Pr\{\mathbf{r}|\mathbf{s}\}$ can be viewed as a normalization constant. Hence,

$$\Pr\{\alpha, \beta|\mathbf{s}, \mathbf{r}\} \propto \prod_i \Pr\{r_i|s_i, \alpha, \beta\}$$

Likelihood function (taking the logarithm and ignoring constants):

$$\begin{aligned}L\left(\alpha, \beta|\mathbf{s}, \mathbf{r}\right) &= \sum_i \log \Pr\{r_i|s_i, \alpha, \beta\} \\ &= \sum_i \begin{cases} \log\left(\Psi\left(\beta\left(s_i - \alpha\right)\right)\right), & \text{for } r_i = \text{TP} \\ \log\left(1 - \Psi\left(\beta\left(s_i - \alpha\right)\right)\right), & \text{for } r_i = \text{FP} \end{cases}\end{aligned}$$

The maximum likelihood estimator is

$$(\alpha^*, \beta^*) = \arg\max_{\alpha,\beta}\left\{L\left(\alpha, \beta|\mathbf{s}, \mathbf{r}\right)\right\}$$

In Figure 8.5, the fitted curve $\Pr\{r = \text{TP}|s\}$ is given for the JSRT database. Figure 8.6a shows the distribution of CAD scores, and Figure 8.6b shows the distribution of probabilities that are associated with those CAD scores. One important remark should be made about the latter: the probabilities for CAD prompts in a certain image are computed by leaving that images out of the computations of the parameters α and β. Otherwise, the probability would be biased.

Figure 8.7 shows two FROC curves, one that is constructed from raw CAD scores and one that is constructed from $\Pr\{r = \text{TP}|s\}$. One could argue that the latter would give better results. If that was true then the reader only needed to copy the given $\Pr\{r = \text{TP}|s\}$ to outperform the CAD system. Luckily this is not possible. Some information seems to be lost due to the conversion of CAD scores, but not much.

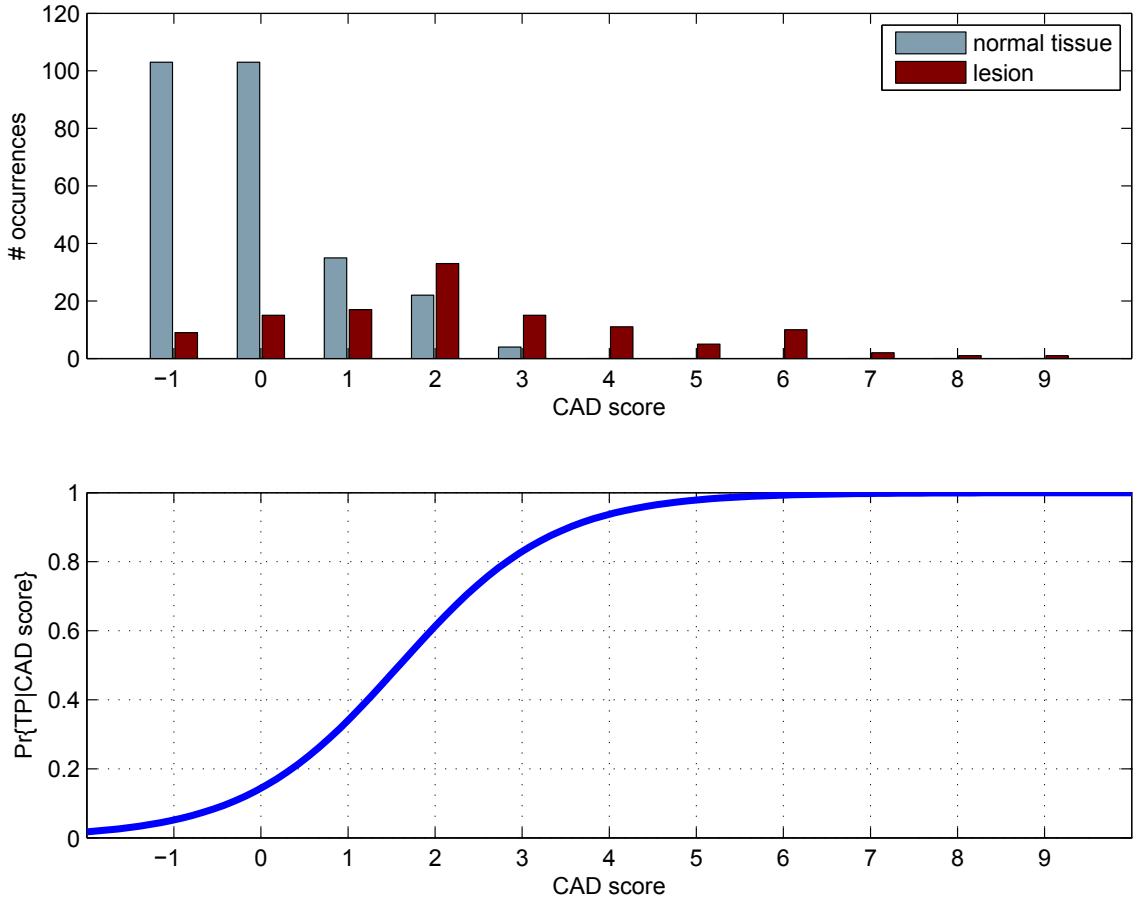
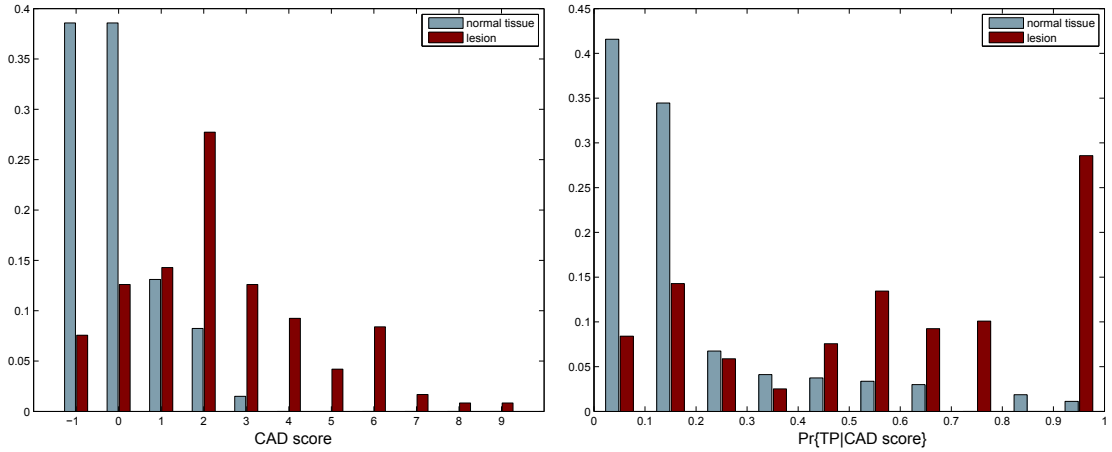


Figure 8.5: The function $\Pr\{r = \text{TP}|s\}$ which has been fitted to all data($\alpha = 1.59$, $\beta = 1.123$)

.



(a) CAD score **(b)** $\Pr\{r = \text{TP}|\text{CAD score}\}$

Figure 8.6: Normalized Histograms. Histograms of CAD scores by Onguard 5.0 for the JSRT database, and the corresponding probabilities $\Pr\{r = \text{TP}|s\}$.

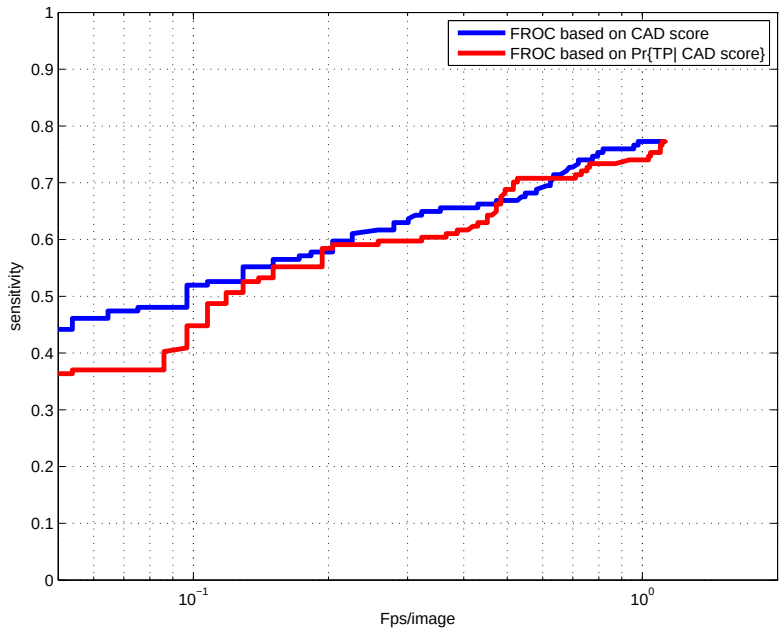


Figure 8.7: FROC analysis. FROC curves constructed from Onguard 5.0 CAD scores and from the corresponding probabilities $\Pr\{\text{TP}|s\}$.

Bibliography

[1] Nodine C. F., Kundel H. L., Lauver S. C., and Toto L. C. Nature of expertise in searching mammograms for breast masses. *Academic Radiology*, 3(12):1000–1006, Dec 1996.

[2] Nodine C. F., Mello-Thoms C., Weinstein S. P., Kundel H. L., Conant E. F., Heller-Savoy R. E., Rowlings S. E., and Birnbaum J. A. Blinded review of retrospectively visible unreported breast cancers: an eye-position analysis. *Radiology*, 221(1):122–129, Oct 2001.

[3] Majid A. S., de Paredes E. S., Doherty R. D., Sharma N. R., and Salvador X. Missed breast carcinoma: pitfalls and pearls. *Radiographics*, 23(4):881–895, 2003.

[4] Manning D. J., Ethell S. C., and Donovan T. Detection or decision errors? Missed lung cancer from the posteroanterior chest radiograph. *British Journal of Radiology*, 77(915):231–5, 2004.

[5] Manning D., Ethell S., and Donovan T. Categories of observer error from eye tracking and AFROC data. In *Proceedings of the SPIE*, volume 5372, pages 90–99, 2004.

[6] Krupinski E. A., Berbaum K. S., Caldwell R. T., Schartz K. M., and Kim J. Long radiology workdays reduce detection and accommodation accuracy. *Journal of the American College of Radiology*, 7 (9):698–704, Sep 2010.

[7] Elmore J. G., Jackson S. L., Abraham L., Miglioretti D. L., Carney P. A., Geller B. M., Yankaskas B. C., Kerlikowske K., Onega T., Rosenberg R. D., Sickles E. A., and Buist D. S. M. Variability in interpretive performance at screening mammography and radiologists' characteristics associated with accuracy. *Radiology*, 253(3):641–651, Dec 2009.

[8] Bird R. E., Wallace T. W., and Yankaskas B. C. Analysis of cancers missed at screening mammography. *Radiology*, 184(3):613–617, 1992.

[9] Mello-Thoms C. How does the perception of a lesion influence visual search strategy in mammogram reading? *Academic Radiology*, 13(3):275–288, Mar 2006.

[10] Kundel H. L. and Nodine C. F. Interpreting chest radiographs without visual search. *Radiology*, 116(3):527–532, Sep 1975.

[11] Kundel H. L. and Revesz G. Lesion conspicuity, structured noise, and film reader error. *American Journal of Roentgenology*, 126(6):1233–1238, Jun 1976.

[12] Kundel H. L., Nodine C. F., and Carmody D. Visual scanning, pattern recognition and decision-making in pulmonary nodule detection. *Investigative Radiology*, 13(3):175–181, 1978.

[13] Lodwick G. S. Computer-aided diagnosis in radiology. a research plan. *Investigative Radiology*, 1 (1):72–80, 1966.

[14] Winsberg F., Elkin M., Macy J., Bordaz V., and weymouth W. Detection of radiographic abnormalities in mammograms by means of optical scanning and computer analysis. *Radiology*, 89:211–215, 1967.

[15] Meyers P. H., Nice Jr. C. M., Becker H. C., Nettleton W. J., Sweeney J. W., and Meckstroth G. R. Automated computer analysis of radiographic images. *Radiology*, 83:1029–1034, 1964.

[16] Roellinger F. X., Kahveci A. E., Chang J. K., Harlow C. A., Dwyer III S. J., and Lodwick G. S. Computer analysis of radiographic images. *Computer Graphics and Image Processing*, 2:232–251, 1973.

[17] Toriwaki J., Suenaga Y., Negoro T., and Fukumura T. Pattern recognition of chest X-ray images.

Computer Graphics and Image Processing, 2:252–271, 1973.

[18] Chan H. P., Doi K., Galhotra S., Vyborny C. J., MacMahon H., and Jokich P. M. Image feature analysis and computer-aided diagnosis in digital radiography. i. automated detection of microcalcifications in mammography. *Medical Physics*, 14(4):538–548, 1987.

[19] Giger M. L., Doi K., and MacMahon H. Image feature analysis and computer-aided diagnosis in digital radiography: automated detection of nodules in peripheral lung fields. *Medical Physics*, 15 (2):158–166, 1988.

[20] Schopper D. and de Wolf C. How effective are breast cancer screening programmes by mammography? review of the current evidence. *European Journal of Cancer*, 45(11):1916–1923, Jul 2009.

[21] Gøtzsche P. C. and Nielsen M. Screening for breast cancer with mammography. *Cochrane Database of Systematic Reviews*, 4:CD001877, 2009.

[22] Burhenne L. J. W., 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, May 2000.

[23] Yankaskas B. C., Schell M. J., Bird R. E., and Desrochers D. A. Reassessment of breast cancers missed during routine screening mammography: a community-based study. *American Journal of Roentgenology*, 177(3):535–541, Sep 2001.

[24] Brem R. F., Baum J., Lechner M., Kaplan S., Souders S., Naul L. G., and Hoffmeister J. Improvement in sensitivity of screening mammography with computer-aided detection: a multiinstitutional trial. *American Journal of Roentgenology*, 181(3):687–693, Sep 2003.

[25] Hofvind S., Skaane P., Vitak B., Wang H., Thoresen S., Eriksen L., Bjrndal H., Braaten A., and Bjurstam N. Influence of review design on percentages of missed interval breast cancers: retrospective study of interval cancers in a population-based screening program. *Radiology*, 237(2): 437–443, Nov 2005.

[26] 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, Sep 1993.

[27] Harvey J. A., Fajardo L. L., and Innis C. A. Previous mammograms in patients with impalpable breast carcinoma: retrospective vs blinded interpretation. 1993 arrs president's award. *American Journal of Roentgenology*, 161(6):1167–1172, Dec 1993.

[28] Martin J. E., Moskowitz M., and Milbrath J. R. Breast cancer missed by mammography. *American Journal of Roentgenology*, 132(5):737–739, May 1979.

[29] Jones R. D., McLean L., Young J. R., Simpson W., and Neilson F. Proportion of cancers detected at the first incident screen which were false negative at the prevalent screen. *The Breast*, 5(5):339–343, 1996.

[30] Bassett L. W., Bunnell D. H., Jahanshahi R., Gold R. H., Arndt R. D., and Linsman J. Breast cancer detection: one versus two views. *Radiology*, 165(1):95–97, Oct 1987.

[31] Vyborny C. J. and Giger M. L. Computer vision and artificial intelligence in mammography. *American Journal of Roentgenology*, 162(3):699–708, Mar 1994.

[32] Tourassi G. D., Delong D. M., and Floyd C. E. A study on the computerized fractal analysis of

architectural distortion in screening mammograms. *Physics in Medicine and Biology*, 51(5):1299–1312, Mar 2006.

[33] Rangayyan R. M., Banik S., and Desautels J. E. L. Computer-aided detection of architectural distortion in prior mammograms of interval cancer. *Journal of Digital Imaging*, 23(5):611–631, Oct 2010.

[34] Ayres F. J. and Rangayyan R. M. Characterization of architectural distortion in mammograms. *IEEE Engineering in Medicine and Biology Magazine*, 24(1):59–67, 2005.

[35] Banik S., Rangayyan R. M., and Desautels J. E. L. Detection of architectural distortion in prior mammograms. *IEEE Transactions on Medical Imaging*, 30(2):279–294, Feb 2011.

[36] Giger M. L., Chan H.-P., and Boone J. Anniversary paper: History and status of cad and quantitative image analysis: the role of medical physics and aapm. *Medical Physics*, 35(12):5799–5820, Dec 2008.

[37] Nishikawa R. M., Giger M. L., Doi K., Metz C. E., Yin F.-F., Vyborny C. J., and Schmidt R. A. Effect of case selection on the performance of computer-aided detection schemes. *Medical Physics*, 21(2): 265–269, 1994.

[38] Zhang W., Doi K., Giger M. L., Wu Y., Nishikawa R. M., and Schmidt R. A. Computerized detection of clustered microcalcifications in digital mammograms using a shift-invariant artificial neural network. *Medical Physics*, 21(4):517–524, Apr 1994.

[39] Nagel R. H., Nishikawa R. M., Papaioannou J., and Doi K. Analysis of methods for reducing false positives in the automated detection of clustered microcalcifications in mammograms. *Medical Physics*, 25(8):1502–1506, Aug 1998.

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

[41] Giger M. L., Vyborny C. J., and Schmidt R. A. Computerized characterization of mammographic masses: analysis of spiculation. *Cancer Letters*, 77(2-3):201–211, Mar 1994.

[42] Yin F. F., Giger M. L., Doi K., Vyborny C. J., and Schmidt R. A. Computerized detection of masses in digital mammograms: automated alignment of breast images and its effect on bilateral-subtraction technique. *Medical Physics*, 21(3):445–452, Mar 1994.

[43] Huo Z., Giger M. L., Vyborny C. J., Bick U., Lu P., Wolverton D. E., and Schmidt R. A. Analysis of spiculation in the computerized classification of mammographic masses. *Medical Physics*, 22(10): 1569–1579, Oct 1995.

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

[45] Sahiner B., Chan H. P., Petrick N., Helvie M. A., and Goodsitt M. M. Computerized characterization of masses on mammograms: the rubber band straightening transform and texture analysis. *Medical Physics*, 25(4):516–526, Apr 1998.

[46] Zheng B., Chang Y. H., Staiger M., Good W., and Gur D. Computer-aided detection of clustered microcalcifications in digitized mammograms. *Academic Radiology*, 2(8):655–662, Aug 1995.

[47] 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, Jul 2001.

[48] 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, May 2010.

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

[50] Zheng B., Chang Y. H., and Gur D. Computerized detection of masses from digitized mammograms: comparison of single-image segmentation and bilateral-image subtraction. *Academic Radiology*, 2(12):1056–1061, Dec 1995.

[51] van Engeland S. and Karssemeijer N. Combining two mammographic projections in a computer aided mass detection method. *Medical Physics*, 34(3):898–905, 2007.

[52] 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, Nov 2009.

[53] 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, Feb 2002.

[54] Wei Q., Hu Y., Gelfand G., and Macgregor J. Segmentation of lung lobes in high-resolution isotropic ct images. *IEEE Transactions on Biomedical Engineering*, 2009.

[55] Timp S., van Engeland S., and Karssemeijer N. A regional registration method to find corresponding mass lesions in temporal mammogram pairs. *Medical Physics*, 32(8):2629–2638, 2005.

[56] 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, Nov 2001.

[57] Sanjay-Gopal S., Chan H. P., Wilson T., Helvie M., Petrick N., and Sahiner B. A regional registration technique for automated interval change analysis of breast lesions on mammograms. *Medical Physics*, 26(12):2669–2679, Dec 1999.

[58] Chan H. P., Doi K., Vyborny C. J., Schmidt R. A., Metz C. E., Lam K. L., Ogura T., Wu Y. Z., and MacMahon H. Improvement in radiologists' detection of clustered microcalcifications on mammograms. the potential of computer-aided diagnosis. *Investigative Radiology*, 25(10):1102–1110, Oct 1990.

[59] Kegelmeyer W. P., Pruneda J. M., Bourland P. D., Hillis A., Riggs M. W., and Nipper M. L. Computer-aided mammographic screening for spiculated lesions. *Radiology*, 191(2):331–337, May 1994.

[60] Karssemeijer N. and Hendriks J. H. Computer-assisted reading of mammograms. *European Radiology*, 7(5):743–748, 1997.

[61] Huo Z., Giger M. L., Vyborny C. J., and Metz C. E. Breast cancer: effectiveness of computer-aided diagnosis observer study with independent database of mammograms. *Radiology*, 224(2):560–568, Aug 2002.

[62] Jiang Y., Nishikawa R. M., Schmidt R. A., and Metz C. E. Comparison of independent double readings and computer-aided diagnosis (cad) for the diagnosis of breast calcifications. *Academic*

Radiology, 13(1):84–94, Jan 2006.

[63] Schmidt R., Nishikawa R., Osnis R., Giger M., Schreibman K., and Doi K. Computerized detection of lesions missed by mammography. In Doi K., Giger M., Nishikawa R., and Schmidt R., editors, *IWDM '96: Proceedings of the 3rd International Workshop on Digital Mammography*, 1119, pages 105–110, Amsterdam, 1996. Excerpta Medica International Congress.

[64] te Brake G. M., Karssemeijer N., and Hendriks J. H. Automated detection of breast carcinomas not detected in a screening program. *Radiology*, 207(2):465–471, 1998.

[65] Philippens M. E. P., Gambarota G., van der Kogel A. J., and Heerschap A. Radiation effects in the rat spinal cord: evaluation with apparent diffusion coefficient versus t2 at serial mr imaging. *Radiology*, 250(2):387–397, Feb 2009.

[66] Gilbert F. J., Astley S. M., Gillan M. G. C., Agbaje O. F., Wallis M. G., James J., Boggis C. R. M., Duffy S. W., and CADET II Group. Single reading with computer-aided detection for screening mammography. *New England Journal of Medicine*, 359:1675–1684, Oct 2008.

[67] Gromet M. Comparison of computer-aided detection to double reading of screening mammograms: review of 231,221 mammograms. *American Journal of Roentgenology*, 190(4):854–859, Apr 2008.

[68] Khoo L. A. L., Taylor P., and Given-Wilson R. M. Computer-aided detection in the United Kingdom National Breast Screening Programme: prospective study. *Radiology*, 237(2):444–449, Nov 2005.

[69] Morton M. J., Whaley D. H., Brandt K. R., and Amrami K. K. Screening mammograms: interpretation with computer-aided detection–prospective evaluation. *Radiology*, 239(2):375–383, May 2006.

[70] Gur D., Sumkin J. H., Rockette H. E., Ganott M., Hakim C., Hardesty L., Poller W. R., Shah R., and Wallace L. Changes in breast cancer detection and mammography recall rates after the introduction of a computer-aided detection system. *Journal of the National Cancer Institute*, 96(3):185–190, Feb 2004.

[71] 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. *New England Journal of Medicine*, 356(14):1399–1409, 2007.

[72] Getty D. J., Pickett R. M., D'Orsi C. J., and Swets J. A. Enhanced interpretation of diagnostic images. *Investigative Radiology*, 23(4):240–252, Apr 1988.

[73] Lo S.-C., Chan H.-P., Lin J.-S., Li H., Freedman M. T., and Mun S. K. Artificial convolution neural network for medical image pattern recognition. *Neural Networks*, 8(7/8):1201–1214, 1995.

[74] Leichter I., Fields S., Nirel R., Bamberger P., Novak B., Lederman R., and Buchbinder S. Improved mammographic interpretation of masses using computer-aided diagnosis. *European Radiology*, 10 (2):377–383, 2000.

[75] Jesneck J. L., Lo J. Y., and Baker J. A. Breast mass lesions: computer-aided diagnosis models with mammographic and sonographic descriptors. *Radiology*, 244(2):390–398, Aug 2007.

[76] Wang Y., Jiang S., Wang H., Guo Y. H., Liu B., Hou Y., Cheng H., and Tian J. Cad algorithms

for solid breast masses discrimination: evaluation of the accuracy and interobserver variability. *Ultrasound in Medicine and Biology*, 36(8):1273–1281, Aug 2010.

[77] Varela C., Timp S., and Karssemeijer N. Use of border information in the classification of mammographic masses. *Physics in Medicine and Biology*, 51(2):425–441, 2006.

[78] Jiang Y., Nishikawa R. M., Schmidt R. A., Metz C. E., Giger M. L., and Doi K. Improving breast cancer diagnosis with computer-aided diagnosis. *Academic Radiology*, 6(1):22–33, Jan 1999.

[79] Chang Y. H., Good W. F., Sumkin J. H., Zheng B., and Gur D. Computerized localization of breast lesions from two views. an experimental comparison of two methods. *Investigative Radiology*, 34 (9):585–588, Sep 1999.

[80] Horsch K., Giger M. L., Vyborny C. J., Lan L., Mendelson E. B., and Hendrick R. E. Classification of breast lesions with multimodality computer-aided diagnosis: observer study results on an independent clinical data set. *Radiology*, 240(2):357–368, Aug 2006.

[81] Sahiner B., Chan H.-P., Hadjiiski L. M., Roubidoux M. A., Paramagul C., Bailey J. E., Nees A. V., Blane C. E., Adler D. D., Patterson S. K., Klein K. A., Pinsky R. W., and Helvie M. A. Multimodality cadx: Roc study of the effect on radiologists' accuracy in characterizing breast masses on mammograms and 3d ultrasound images. *Academic Radiology*, 16(7):810–818, Jul 2009.

[82] Drukker K., Horsch K., and Giger M. L. Multimodality computerized diagnosis of breast lesions using mammography and sonography. *Academic Radiology*, 12(8):970–979, Aug 2005.

[83] 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.

[84] Blanks R. G., Wallis M. G., and Given-Wilson R. M. Observer variability in cancer detection during routine repeat (incident) mammographic screening in a study of two versus one view mammography. *Journal of Medical Screening*, 6(3):152–158, 1999.

[85] Karssemeijer N., Otten J. D. M., Rijken H., and Holland R. Computer aided detection of masses in mammograms as decision support. *British Journal of Radiology*, 79 Spec No 2:S123–S126, 2006.

[86] 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, Apr 2008.

[87] 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.

[88] Samulski M., Hupse R., Boetes C., Mus R., den Heeten G., and Karssemeijer N. Using Computer Aided Detection in Mammography as a Decision Support. *European Radiology*, 20(10):2323–2330, June 2010.

[89] Swett H. A., Fisher P. R., Cohn A. I., Miller P. L., and Mutalik P. G. Expert system-controlled image display. *Radiology*, 172(2):487–493, Aug 1989.

[90] Tourassi G. D., Vargas-Voracek R., Catarious D. M., and Floyd C. E. Computer-assisted detection of mammographic masses: a template matching scheme based on mutual information. *Medical Physics*, 30(8):2123–2130, Aug 2003.

[91] Tourassi G. D., Harrawood B., Singh S., Lo J. Y., and Floyd C. E. Evaluation of information-

theoretic similarity measures for content-based retrieval and detection of masses in mammograms. *Medical Physics*, 34(1):140–150, Jan 2007.

[92] Zheng B., Mello-Thoms C., Wang X.-H., Abrams G. S., Sumkin J. H., Chough D. M., Ganott M. A., Lu A., and Gur D. Interactive computer-aided diagnosis of breast masses: computerized selection of visually similar image sets from a reference library. *Academic Radiology*, 14(8):917–927, Aug 2007.

[93] Horsch K., Giger M. L., and Metz C. E. Prevalence scaling: applications to an intelligent workstation for the diagnosis of breast cancer. *Academic Radiology*, 15(11):1446–1457, Nov 2008.

[94] Mettler F. A., Bhargavan M., Faulkner K., Gilley D. B., Gray J. E., Ibbott G. S., Lipoti J. A., Mahesh M., McCrohan J. L., Stabin M. G., Thomadsen B. R., and Yoshizumi T. T. Radiologic and nuclear medicine studies in the United States and worldwide: frequency, radiation dose, and comparison with other radiation sources–1950-2007. *Radiology*, 253(2):520–531, Nov 2009.

[95] Monnier-Cholley L., Arrivé L., Porcel A., Shehata K., Dahan H., Urban T., Febvre M., Lebeau B., and Tubiana J. M. Characteristics of missed lung cancer on chest radiographs: a French experience. *European Radiology*, 11(4):597–605, 2001.

[96] Shah P. K., Austin J. H. M., White C. S., Patel P., Haramati L. B., Pearson G. D. N., Shiau M. C., and Berkmen Y. M. Missed non-small cell lung cancer: Radiographic findings of potentially resectable lesions evident only in retrospect. *Radiology*, 226(1):235–241, 2003.

[97] Muhm J. R., Miller W. E., Fontana R. S., Sanderson D. R., and Uhlenhopp M. A. Lung cancer detected during a screening program using four-month chest radiographs. *Radiology*, 148:609–615, 1983.

[98] Quekel L. G., Kessels A. G., Goei R., and Engelshoven J. M. v. Miss rate of lung cancer on the chest radiograph in clinical practice. *Chest*, 115(3):720–724, 1999.

[99] Gohagan J. K., Marcus P. M., Fagerstrom R. M., Pinsky P. F., Kramer B. S., Prorok P. C., Ascher S., Bailey W., Brewer B., Church T., Engelhard D., Ford M., Fouad M., Freedman M., Gelmann E., Gierada D., Hocking W., Inampudi S., Irons B., Johnson C. C., Jones A., Kucera G., Kvale P., Lappe K., Manor W., Moore A., Nath H., Neff S., Oken M., Moore A., Plunkett M., Price H., Reding D., Riley T., Schwartz M., Spizarny D., Yoffie R., Zylak C., and the Lung Screening Study Research Group. Final results of the lung screening study, a randomized feasibility study of spiral ct versus chest x-ray screening for lung cancer. *Lung Cancer*, 47(1):9–15, 2005.

[100] Giger M. L., Doi K., MacMahon H., and Metz C. E. Computerized detection of pulmonary nodules in digital chest images: use of morphological filters in reducing false positive detections. *Medical Physics*, 17(5):861–865, 1990.

[101] Abe H., Macmahon H., Engelmann R., Li Q., Shiraishi J., Katsuragawa S., Aoyama M., Ishida T., Ashizawa K., e. Metz C., and Doi K. Computer-aided diagnosis in chest radiography: Results of large-scale observer tests at the 1996–2001 RSNA scientific assemblies. *Radiographics*, 23:255–265, 2003.

[102] White C. S., Flukinger T., Jeudy J., and Chen J. J. Use of a computer-aided detection system to detect missed lung cancer at chest radiography. *Radiology*, 252(1):273–281, Jul 2009.

[103] de Boo D. W., Prokop M., Uffmann M., van Ginneken B., and Schaefer-Prokop C. M. Computer-aided detection (CAD) of lung nodules and small tumours on chest radiographs. *European Journal*

of Radiology, 72(2):218–225, 2009.

[104] Hardie R. C., Rogers S. K., Wilson T., and Rogers A. Performance analysis of a new computer aided detection system for identifying lung nodules on chest radiographs. *Medical Image Analysis*, 12(3):240–58, 2008.

[105] Schilham A. M. R. and van Ginneken B. Simulating nodules in chest radiographs with real nodules from multi-slice CT images. In *Medical Imaging*, number 6144 in Proceedings of the SPIE, pages 614456–1–614456–8, 2006.

[106] de Hoop B., Gietema H., van de Vorst S., Murphy K., van Klaveren R. J., and Prokop M. Pulmonary ground-glass nodules: increase in mass as an early indicator of growth. *Radiology*, 255(1):199–206, Apr 2010.

[107] van Beek E. J. R., Mullan B., and Thompson B. Evaluation of a real-time interactive pulmonary nodule analysis system on chest digital radiographic images: a prospective study. *Academic Radiology*, 15(5):571–5, 2008.

[108] Kasai S., Li F., Shiraishi J., and Doi K. Usefulness of computer-aided diagnosis schemes for vertebral fractures and lung nodules on chest radiographs. *American Journal of Roentgenology*, 191(1): 260–5, 2008.

[109] Kobayashi T., Xu X.-W., MacMahon H., Metz C., and Doi K. Effect of a computer-aided diagnosis scheme on radiologists' performance in detection of lung nodules on radiographs. *Radiology*, 199: 843–848, 1996.

[110] Awai K., Murao K., Ozawa A., Komi M., Hayakawa H., Hori S., and Nishimura Y. Pulmonary nodules at chest CT: Effect of computer-aided diagnosis on radiologists' detection performance. *Radiology*, 230(2):347–352, 2004.

[111] Monnier-Cholley L., Carrat F., Cholley B. P., Tubiana J., and Arrivé L. Detection of lung cancer on radiographs: receiver operating characteristic analyses of radiologists', pulmonologists', and anesthesiologists' performance. *Radiology*, 233(3):799–805, 2004.

[112] Das M., Mühlenbruch G., Mahnken A. H., Flohr T. G., Gündel L., Stanzel S., Kraus T., Günther R. W., and Wildberger J. E. Small pulmonary nodules: effect of two computer-aided detection systems on radiologist performance. *Radiology*, 241(2):564–571, 2006.

[113] Armato S. G., Giger M. L., and MacMahon H. Automated detection of lung nodules in CT scans: Preliminary results. *Medical Physics*, 28(8):1552–1561, 2001.

[114] Suzuki K., Armato S. G., Li F., Sone S., and Doi K. Massive training artificial neural network (MTANN) for reduction of false positives in computerized detection of lung nodules in low-dose computed tomography. *Medical Physics*, 30(7):1602–1617, 2003.

[115] Rubin G. D., Lyo J. K., Paik D. S., Sherbondy A. J., Chow L. C., Leung A. N., Mindelzun R., Schraedley-Desmond P. K., Zinck S. E., Naidich D. P., and Napel S. Pulmonary nodules on multi?detector row CT scans: Performance comparison of radiologists and computer-aided detection. *Radiology*, 234:274–283, 2005.

[116] Sahiner B., Chan H.-P., Hadjiiski L. M., Cascade P. N., Kazerooni E. A., Chughtai A. R., Poopat C., Song T., Frank L., Stojanovska J., and Attili A. Effect of cad on radiologists' detection of lung nodules on thoracic ct scans: analysis of an observer performance study by nodule size. *Academic*

Radiology, 16(12):1518–1530, Dec 2009.

[117] Roos J. E., Paik D., Olsen D., Liu E. G., Chow L. C., Leung A. N., Mindelzun R., Choudhury K. R., Naidich D. P., Napel S., and Rubin G. D. Computer-aided detection (cad) of lung nodules in ct scans: radiologist performance and reading time with incremental cad assistance. *European Radiology*, 20(3):549–557, Mar 2010.

[118] Way T., Chan H.-P., Hadjiiski L., Sahiner B., Chughtai A., Song T. K., Poopat C., Stojanovska J., Frank L., Attili A., Bogot N., Cascade P. N., and Kazerooni E. A. Computer-aided diagnosis of lung nodules on CT scans: ROC study of its effect on radiologists' performance. *Academic Radiology*, 17 (3):323–332, Mar 2010.

[119] MacMahon H., Engelmann R., Behlen F. M., Hoffmann K. R., Ishida T., Roe C., Metz C. E., and Doi K. Computer-aided diagnosis of pulmonary nodules: results of a large-scale observer test. *Radiology*, 213:723–726, 1999.

[120] Ashizawa K., MacMahon H., Ishida T., Nakamura K., Vyborny C. J., Katsuragawa S., and Doi K. Effect of an artificial neural network on radiologists' performance in the differential diagnosis of interstitial lung disease using chest radiographs. *American Journal of Roentgenology*, 172:1311–1315, 1999.

[121] Shiraishi J., Abe H., Engelmann R., Aoyama M., MacMahon H., and Doi K. Computer-aided diagnosis to distinguish benign from malignant solitary pulmonary nodules on radiographs: Roc analysis of radiologists' performance–initial experience. *Radiology*, 227(2):469–474, 2003.

[122] Summers R. M., Liu J., Rehani B., Stafford P., Brown L., Louie A., Barlow D. S., Jensen D. W., Cash B., Choi J. R., Pickhardt P. J., and Petrick N. Ct colonography computer-aided polyp detection: Effect on radiologist observers of polyp identification by cad on both the supine and prone scans. *Academic Radiology*, 17(8):948–959, Aug 2010.

[123] Baker M. E., Bogoni L., Obuchowski N. A., Dass C., Kendzierski R. M., Remer E. M., Einstein D. M., Cathier P., Jerebko A., Lakare S., Blum A., Caroline D. F., and Macari M. Computer-aided detection of colorectal polyps: can it improve sensitivity of less-experienced readers? preliminary findings. *Radiology*, 245(1):140–149, Oct 2007.

[124] Metz C. E. Roc methodology in radiologic imaging. *Investigative Radiology*, 21(9):720–733, 1986.

[125] Starr S. J., Metz C. E., Lusted L. B., and Goodenough D. J. Visual detection and localization of radiographic images. *Radiology*, 116(3):533–538, Sep 1975.

[126] Bunch P., Hamilton J., Sanderson G., and Simmons A. A free response approach to the measurement and characterization of radiographic-observer performance. *Journal of Applied Photographic Engineering*, 4:166–172, 1978.

[127] Chakraborty D. P. and Winter L. H. Free-response methodology: alternate analysis and a new observer-performance experiment. *Radiology*, 174(3 Pt 1):873–881, Mar 1990.

[128] Metz C. Rockit v1.1b2 receiver operating characteristic software, 2006. URL `http://metz-roc.uchicago.edu/`.

[129] Metz C. Dbm mrmc - multi-reader, multi-case stastical analysis software, 2008. URL `http://metz-roc.uchicago.edu/`.

[130] Chakraborty D. Statistical power in observer-performance studies: comparison of the receiver

operating characteristic and free-response methods in tasks involving localization. *Academic Radiology*, 9(2):147–156, Feb 2002.

[131] Edwards D. C., Kupinski M. A., Metz C. E., and Nishikawa R. M. Maximum likelihood fitting of froc curves under an initial-detection-and-candidate-analysis model. *Medical Physics*, 29(12): 2861–2870, Dec 2002.

[132] Chakraborty D. P. and Berbaum K. S. Observer studies involving detection and localization: modeling, analysis, and validation. *Medical Physics*, 31(8):2313–2330, Aug 2004.

[133] Chakraborty D. P. Analysis of location specific observer performance data: validated extensions of the jackknife free-response (jafroc) method. *Academic Radiology*, 13(10):1187–1193, Oct 2006.

[134] Samuelson F. and Petrick N. Comparing image detection algorithms using resampling. In *Biomedical Imaging: Nano to Macro, 2006. 3rd IEEE International Symposium on*, pages 1312 –1315, 6-9 2006.

[135] Samuelson F., Petrick N., and Paquerault S. Advantages and examples of resampling for CAD evaluation. In *Biomedical Imaging: From Nano to Macro, 2007. ISBI 2007. 4th IEEE International Symposium on*, pages 492 –495, 12-15 2007.

[136] Hupse R. and Karssemeijer N. The use of contextual information for computer aided detection of masses in mammograms. In *Medical Imaging*, volume 7260 of *Proceedings of the SPIE*, page 72600Q, 2009.

[137] Chakraborty D. P. and Yoon H.-J. JAFROC analysis revisited: figure-of-merit considerations for human observer studies. In Sahiner B. and Manning D. J., editors, *Medical Imaging*, volume 7263 of *Proceedings of the SPIE*, page 72630T, 2009.

[138] Burnside E., Rubin D., and Shachter R. A Bayesian network for mammography. In *Proceedings of AMIA Annual Symposium*, pages 106–110, 2000.

[139] Wang X. H., Zheng B., Good W. F., King J. L., and Chang Y. H. Computer-assisted diagnosis of breast cancer using a data-driven bayesian belief network. *International Journal of Medical Informatics*, 54(2):115–126, May 1999.

[140] Cortes C. and Vapnik V. Support-vector networks. *Machine Learning*, 20(3):273–97, 1995.

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

[142] Nattkemper T. W., Arnrich B., Lichte O., Timm W., Degenhard A., Pointon L., Hayes C., Leach M. O., and Study U. K. M. B. S. Evaluation of radiological features for breast tumour classification in clinical screening with machine learning methods. *Artificial Intelligence Medical*, 34(2):129–139, Jun 2005.

[143] Mavroforakis M., Georgiou H., Dimitropoulos N., Cavouras D., and Theodoridis S. Significance analysis of qualitative mammographic features, using linear classifiers, neural networks and support vector machines. *European Journal of Radiology*, 54(1):80–89, Apr 2005.

[144] Li S., Fevens T., Krzyzak A., and Li S. An automatic variational level set segmentation framework for computer aided dental x-rays analysis in clinical environments. *Comput Med Imaging Graph*, 30(2):65–74, Mar 2006.

[145] te Brake G. M., Karssemeijer N., and Hendriks J. H. An automatic method to discriminate malignant masses from normal tissue in digital mammograms. *Physics in Medicine and Biology*, 45(10):

2843–2857, 2000.

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

[147] Snoeren P. R. and Karssemeijer N. Thickness correction of mammographic images by anisotropic filtering and interpolation of dense tissue. In *Medical Imaging*, volume 5747 of *Proceedings of the SPIE*, pages 1521–1527, 2005.

[148] Timp S. and Karssemeijer N. A new 2D segmentation method based on dynamic programming applied to computer aided detection in mammography. *Medical Physics*, 31(5):958–971, 2004.

[149] Timp S. *Analysis of Temporal Mammogram Pairs to Detect and Characterise Mass Lesions*. PhD thesis, Radboud University Nijmegen Medical Centre, 2006.

[150] Caulkin S., Astley S., Asquith J., and Boggis C. Sites of occurrence of malignancies in mammograms. In *IWDM '98: Proceedings of the 4th international workshop on Digital Mammography*, volume 13, pages 279–282. Springer Berlin / Heidelberg, 1998.

[151] Metz C. E., Herman B. A., and Shen J. H. Maximum likelihood estimation of receiver operating characteristic (roc) curves from continuously-distributed data. *Statistics in Medicine*, 17(9):1033–1053, May 1998.

[152] Metz C., Wang P., and Kronman H. A new approach for testing the significance for differences between ROC curves measured from correlated data. *Information Processing in Medical Imaging*, 1984.

[153] Neapolitan R. *Learning Bayesian Networks*. Prentice Hall, Upper Saddle River, NJ, 2003.

[154] Domingos P. and Pazzani M. J. On the optimality of the simple bayesian classifier under zero-one loss. *Machine Learning*, 29(2-3):103–130, 1997.

[155] Smola A. J. and Scholkopf B. A tutorial on support vector regression. *Statistics and Computing*, 14 (3):199–222, 2004.

[156] Murphy K. The bayes net toolbox for Matlab, 2001. URL `http://bnt.sourceforge.net/`.

[157] Duda R. O., Hart P. E., and Stork D. G. *Pattern classification*. John Wiley and Sons, New York, 2nd edition, 2001.

[158] IARC (International Agency for Research on Cancer). Breast cancer and screening. Technical report, World Health Organization. URL `http://www.emro.who.int/ncd/publications/breastcancerscreening.pdf`. accessed on 25-02-2008.

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

[160] van Engeland S., Timp S., and Karssemeijer N. Finding corresponding regions of interest in mediolateral oblique and craniocaudal mammographic views. *Medical Physics*, 33(9):3203–3212, 2006.

[161] Sun X., Qian W., Song D., and Clark R. Ipsilateral multi-view CAD system for mass detection in digital mammography. In *Proceedings of IEEE Workshop on Mathematical Methods in Biomedical Image Analysis*. IEEE Computer Society, 2001.

[162] Qian W., Song D., Lei M., Sankar R., and Eikman E. Computer-aided mass detection based on

ipsilateral multiview mammograms. *Academic Radiology*, 14(5):530–538, 2007.

[163] Jensen F. and Nielsen T. *Bayesian Networks and Decision Graphs*. Springer Verlag, 2007.

[164] Heckerman D. and Breese J. S. Causal independence for probability assessment and inference using Bayesian networks. *IEEE Transactions on Systems, Man and Cybernetics, Part A*, 26(6):826–831, 1996.

[165] Lucas P. J. F. Bayesian network modelling through qualitative pattern. *Artificial Intelligence*, 163: 233–263, 2005.

[166] Diez F. Parameter adjustment in Bayes networks: The generalized noisy or-gate. In *Proceedings of the Ninth Conference on UAI, San Francisco, CA*. Morgan Kaufmann, 1993.

[167] Sarkar S. and Boyer K. Integration, inference, and management of spatial information using bayesian networks: Perceptual organization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(3):256–274, 1993.

[168] Dempster A., Laird N., and Rubin D. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38, 1977.

[169] Hanley J. A. and McNeil B. J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36, 1982.

[170] Kalman B., Reinus W., and Kwasny S. Prescreening entire mammograms for masses with artificial neural networks: Preliminary results. *Academic Radiology*, 4:405–414, 1997.

[171] Astley S., Mistry T., Boggis C., and Hiller V. Should we use humans or a machine to pre-screen mammograms? In H-O P., editor, *Proceedings of Sixth International Workshop in Digital Mammography*, pages 476–480. Springer, 2002.

[172] Moberg K., Bjurstam N., Wilczek B., Rostgard L., Egge E., and Muren C. Computed assisted detection of interval breast cancers. *European Journal of Radiology*, 39(2):104–110, 2001.

[173] Highnam R., Kita Y., Brady M., Shepstone B., and English R. Determining correspondence between views. In Karssemeijer N., Thijssen M., Hendriks J., and van Erning L., editors, *4th International Workshop on Digital Mammography*, pages 111–118, 1998.

[174] Bishop C. M. *Pattern Recognition and Machine Learning*. Springer, New York, 2006. ISBN 0387310738.

[175] Dudani S. The distance-weighted k-nearest-neighbor rule. *IEEE Transactions on Systems, Man, and Cybernetics*, 6(4):325–327, April 1976.

[176] Yuan Y., Giger M. L., Li H., and Sennett C. Correlative feature analysis on FFDM. *Medical Physics*, 35(12):5490–5500, 2008.

[177] Wei J., Chan H.-P., Sahiner B., Zhou C., Hadjiiski L. M., Roubidoux M. A., and Helvie M. A. Computer-aided detection of breast masses on mammograms: dual system approach with two-view analysis. *Medical Physics*, 36(10):4451–4460, Oct 2009.

[178] Velikova M., Samulski M., Lucas P. J. F., and Karssemeijer N. Improved mammographic CAD performance using multi-view information: a Bayesian network framework. *Physics in Medicine and Biology*, 54(5):1131–1147, 2009.

[179] Zheng B., Leader J. K., Abrams G. S., Lu A. H., Wallace L. P., Maitz G. S., and Gur D. Multiview-

based computer-aided detection scheme for breast masses. *Medical Physics*, 33(9):3135–3143, Sep 2006.

[180] te Brake G. M. and Karssemeijer N. Single and multiscale detection of masses in digital mammograms. *IEEE Transactions on Medical Imaging*, 18(7):628–639, 1999.

[181] Altrichter M., Ludányi Z., and Horváth G. Joint Analysis of Multiple Mammographic Views in CAD Systems for Breast Cancer Detection. In *Image Analysis*, volume 3540 of *Lecture Notes in Computer Science*, pages 760–769, 2005.

[182] 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.

[183] Kopans D. B. *Breast Imaging*. Lippincott Williams & Wilkins, 3rd edition edition, 2006.

[184] Altrichter M. and Horváth G. The refinement of microcalcification cluster assessment by joint analysis of mlo and cc views. In Astley S., Brady M., Rose C., and Zwiggelaar R., editors, *Digital Mammography*, volume 4046 of *Lecture Notes in Computer Science*, pages 509–516. Springer Berlin / Heidelberg, 2006.

[185] Samulski M. and Karssemeijer N. Matching mammographic regions in mediolateral oblique and cranio caudal views: a probabilistic approach. In Giger M. L. and Karssemeijer N., editors, *Medical Imaging*, volume 6915 of *Proceedings of the SPIE*, page 69151M, 2008.

[186] Pudil P., Novovicova J., and Kittler J. Floating search methods in feature selection. *Pattern Recognition Letters*, 15(11):1119–1125, 1994.

[187] Spence C. D. and Sajda P. Role of feature selection in building pattern recognizers for computer-aided diagnosis. In Hanson K. M., editor, *Medical Imaging*, volume 3338 of *Proceedings of the SPIE*, pages 1434–1441, 1998.

[188] Dean J. C. and Ilvento C. C. Improved cancer detection using computer-aided detection with diagnostic and screening mammography: prospective study of 104 cancers. *American Journal of Roentgenology*, 187(1):20–28, Jul 2006.

[189] Gur D., Stalder J. S., Hardesty L. A., Zheng B., Sumkin J. H., Chough D. M., Shindel B. E., and Rockette H. E. Computer-aided detection performance in mammographic examination of masses: assessment. *Radiology*, 233(2):418–423, Nov 2004.

[190] Houssami N., Given-Wilson R., and Ciatto S. Early detection of breast cancer: overview of the evidence on computer-aided detection in mammography screening. *Journal of Medical Imaging and Radiation Oncology*, 53(2):171–176, Apr 2009.

[191] Alberdi E., Povyakalo A., Strigini L., Ayton P., and Given-Wilson R. CAD in mammography: lesion-level versus case-level analysis of the effects of prompts on human decisions. *International Journal of Computer Assisted Radiology and Surgery*, 3(1):115–122, 2008.

[192] Taplin S. H., Rutter C. M., and Lehman C. D. Testing the effect of computer-assisted detection on interpretive performance in screening mammography. *American Journal of Roentgenology*, 187(6): 1475–1482, Dec 2006.

[193] Garvican L. and Field S. A pilot evaluation of the R2 image checker system and users' response in the detection of interval breast cancers on previous screening films. *Clinical Radiology*, 56(10): 833–837, Oct 2001.

[194] Alberdi E., Povyakalo A. A., Strigini L., Ayton P., Hartswood M., Procter R., and Slack R. Use of computer-aided detection (CAD) tools in screening mammography: a multidisciplinary investigation. *British Journal of Radiology*, 78 Spec No 1:S31–S40, 2005.

[195] Taylor P., Given-Wilson R., Champness J., Potts H. W. W., and Johnston K. Assessing the impact of CAD on the sensitivity and specificity of film readers. *Clinical Radiology*, 59(12):1099–1105, Dec 2004.

[196] Gilbert F. J., Astley S. M., Boggis C. R., McGee M. A., Griffiths P. M., Duffy S. W., Agbaje O. F., Gillan M. G., Wilson M., Jain A. K., Barr N., Beetles U. M., Griffiths M. A., Johnson J., Roberts R. M., Deans H. E., Duncan K. A., and Iyengar G. Variable size computer-aided detection prompts and mammography film reader decisions. *Breast Cancer Research*, 10(4):R72, 2008.

[197] Karssemeijer N., Hupse A., Samulski M., Kallenberg M., Boetes C., and Heeten G. An Interactive Computer Aided Decision Support System for Detection of Masses in Mammograms. In *IWDM '08: Proceedings of the 9th international workshop on Digital Mammography*, pages 273–278, Berlin, Heidelberg, 2008. Springer-Verlag. ISBN 978-3-540-70537-6.

[198] Taylor P., Champness J., Given-Wilson R., Johnston K., and Potts H. Impact of computer-aided detection prompts on the sensitivity and specificity of screening mammography. *Health Technol Assess*, 9(6):iii, 1–iii,58, Feb 2005.

[199] Roelofs A., van Woudenberg S., Hendriks J., Evertsz C., and Karssemeijer N. Effects of computer-aided diagnosis on radiologists detection of breast masses. In Pisano E., editor, *IWDM '04: Proceedings of the 7th International Workshop on Digital Mammography*, pages 219–224, Chapel Hill, NC, 2004.

[200] Kamitani T., Yabuuchi H., Soeda H., Matsuo Y., Okafuji T., Sakai S., Furuya A., Hatakenaka M., Ishii N., and Honda H. Detection of masses and microcalcifications of breast cancer on digital mammograms: comparison among hard-copy film, 3-megapixel liquid crystal display (LCD) monitors and 5-megapixel LCD monitors: an observer performance study. *European Radiology*, 17 (5):1365–1371, May 2007.

[201] Gur D., Bandos A. I., Cohen C. S., Hakim C. M., Hardesty L. A., Ganott M. A., Perrin R. L., Poller W. R., Shah R., Sumkin J. H., Wallace L. P., and Rockette H. E. The "laboratory" effect: comparing radiologists' performance and variability during prospective clinical and laboratory mammography interpretations. *Radiology*, 249(1):47–53, Oct 2008.

[202] Mello-Thoms C. Perception of breast cancer: eye-position analysis of mammogram interpretation. *Academic Radiology*, 10(1):4–12, Jan 2003.

[203] 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, Nov 2009.

[204] Samulski M., Snoeren P., Platel B., van Ginneken B., Hogeweg L., Schaefer-Prokop C., and Karssemeijer N. Computer-Aided Detection as a Decision Assistant in Chest Radiography. In *Medical Imaging*, volume 7966 of *Proceedings of the SPIE*, page 796614, 2011.

[205] Kallenberg M. and Karssemeijer N. Computer-aided detection of masses in full-field digital mammography using screen-film mammograms for training. *Physics in Medicine and Biology*, 53(23):

6879–6891, Dec 2008.

[206] Wei Q., Hu Y., Gelfand G., and Macgregor J. H. Segmentation of lung lobes in isotropic ct images using wavelet transformation. In *Conf Proc IEEE Eng Med Biol Soc*, volume 1, pages 5551–5554, 2007.

[207] Chakraborty D. P. Maximum likelihood analysis of free-response receiver operating characteristic (froc) data. *Medical Physics*, 16(4):561–568, 1989.

[208] Dorfman D. D., Berbaum K. S., and Metz C. E. Receiver operating characteristic rating analysis: Generalization to the population of readers and patients with the jackknife method. *Investigative Radiology*, 27:723–731, 1992.

[209] Astley S. M. Evaluation of computer-aided detection (cad) prompting techniques for mammography. *British Journal of Radiology*, 78 Spec No 1:S20–S25, 2005.

[210] 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, Oct 2010.

[211] 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.

[212] Hadjiiski L., Sahiner B., Helvie M. A., Chan H.-P., Roubidoux M. A., Paramagul C., Blane C., Petrick N., Bailey J., Klein K., Foster M., Patterson S. K., Adler D., Nees A. V., and Shen J. Breast masses: computer-aided diagnosis with serial mammograms. *Radiology*, 240(2):343–356, Aug 2006.

[213] Chan H. P., Sahiner B., Helvie M. A., Petrick N., Roubidoux M. A., Wilson T. E., Adler D. D., Paramagul C., Newman J. S., and Sanjay-Gopal S. Improvement of radiologists' characterization of mammographic masses by using computer-aided diagnosis: an roc study. *Radiology*, 212(3): 817–827, Sep 1999.

[214] Sahiner B., Chan H.-P., Roubidoux M. A., Hadjiiski L. M., Helvie M. A., Paramagul C., Bailey J., Nees A. V., and Blane C. Malignant and benign breast masses on 3d us volumetric images: effect of computer-aided diagnosis on radiologist accuracy. *Radiology*, 242(3):716–724, Mar 2007.

[215] 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*, 127:2893–2917, Jun 2010.

[216] Berbaum K. S., Franken E. A., Dorfman D. D., Miller E. M., Caldwell R. T., Kuehn D. M., and Berbaum M. L. Role of faulty visual search in the satisfaction of search effect in chest radiography. *Academic Radiology*, 5(1):9–19, Jan 1998.

[217] de Hoop B., de Boo D. W., Gietema H. A., van Hoorn F., Mearadji B., Schijf L., van Ginneken B., Prokop M., and Schaefer-Prokop C. Computer-aided detection of lung cancer on chest radiographs: Effect on observer performance. *Radiology*, 257(2):532–540, 2010.

[218] Shiraishi J., Katsuragawa S., Ikezoe J., Matsumoto T., Kobayashi T., Komatsu K., Matsui M., Fujita H., Kodera Y., and Doi K. Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists' detection of

pulmonary nodules. *American Journal of Roentgenology*, 174:71–74, 2000.

Publications

Papers in international journals

M. Samulski, P. Snoeren, C. Schaefer-Prokop, B. van Ginneken and N. Karssemeijer. "A Novel Method for Presentation of Computer-aided Detection Results in Chest Radiographs", *To be submitted*.

A. Hupse, **M. Samulski**, C. Boetes, R. Mus, G. den Heeten and N. Karssemeijer. "Computer aided detection of masses in mammography: interactive decision support versus prompting", *Submitted to Radiology*.

G. van Schie, C. Tanner, P. Snoeren, **M. Samulski**, K. Leifland, M.G. Wallis and N. Karssemeijer. "Correlating locations in ipsilateral breast tomosynthesis views using an analytical hemispherical compression model", Physics in Medicine and Biology 2011;56(15):4715–4730.

M. Samulski and N. Karssemeijer. "Optimizing Case-based Detection Performance in a Multiview CAD System for Mammography", IEEE Transactions on Medical Imaging 2011;30(4):1001–1009.

M. Samulski, R. Hupse, C. Boetes, R. Mus, G. den Heeten and N. Karssemeijer. "Using Computer Aided Detection in Mammography as a Decision Support", European Radiology 2010;20(10):2323–2330.

M. Velikova, **M. Samulski**, P.J.F. Lucas and N. Karssemeijer. "Improved mammographic CAD performance using multi-view information: a Bayesian network framework", Physics in Medicine and Biology 2009;54(5):1131–1147.

Papers in conference proceedings

M. Samulski, P. Snoeren, B. Platel, B. van Ginneken, L. Hogeweg, C. Schaefer-Prokop and N. Karssemeijer. "Computer-Aided Detection as a Decision Assistant in Chest Radiography", in: Medical Imaging, volume 7966 of Proceedings of the SPIE, 2011, page 796614.

J. Lesniak, R. Hupse, M. Kallenberg, **M. Samulski**, R. Blanc, N. Karssemeijer and G. Szekely. "Computer Aided Detection of Breast Masses in Mammography using Support Vector Machine Classification", in: Medical Imaging, volume 7963 of Proceedings of the SPIE, 2011, page 79631K.

S. Robben, M. Velikova, P.J. Lucas and **M. Samulski**. "Discretisation Does Affect the

Performance of Bayesian Networks", in: Research and Development in Intelligent Systems XXVII, 2011, pages 237–250.

M. Radstake, M. Velikova, P. Lucas and **M. Samulski**. "Critiquing Knowledge Representation in Medical Image Interpretation using Structure Learning", in: Knowledge Representation for Health-Care (KR4HC), volume 6512 of Lecture Notes in Artificial Intelligence, 2011, pages 56–70.

M. Samulski, A. Hupse, C. Boetes, G. den Heeten and N. Karssemeijer. "Analysis of probed regions in an interactive CAD system for the detection of masses in mammograms", in: Medical Imaging, volume 7263 of Proceedings of the SPIE, 2009, page 1, page 726314.

M. Velikova, **M. Samulski**, P.J. Lucas and N. Karssemeijer. "Causal Probabilistic Modelling for Two-View Mammographic Analysis", in: AIME '09: Proceedings of the 12th Conference on Artificial Intelligence in Medicine, 2009, pages 395–404.

N. Karssemeijer, A. Hupse, **M. Samulski**, M. Kallenberg, C. Boetes and G. Heeten. "An Interactive Computer Aided Decision Support System for Detection of Masses in Mammograms", in: IWDM '08: Proceedings of the 9th international workshop on Digital Mammography, 2008, pages 273–278.

M. Samulski and N. Karssemeijer. "Linking mass regions in mediolateral oblique and cranio caudal views", in: Proceedings of the 14th ASCI conference, 2008, pages 214–221.

M. Samulski and N. Karssemeijer. "Matching mammographic regions in mediolateral oblique and cranio caudal views: a probabilistic approach", in: Medical Imaging, volume 6915 of Proceedings of the SPIE, 2008, page 1, page 69151M.

M. Velikova, H. Daniels and **M. Samulski**. "Partially Monotone Networks Applied to Breast Cancer Detection on Mammograms", in: ICANN '08: Proceedings of the 18th international conference on Artificial Neural Networks, Part I, 2008, pages 917–926.

M. Velikova, P.J.F. Lucas, N. Ferreira, **M. Samulski** and N. Karssemeijer. "A decision support system for breast cancer detection in screening programs", in: Proceeding of the 2008 conference on ECAI 2008, 2008, pages 658–662.

M. Velikova, **M. Samulski**, N. Karssemeijer and P. Lucas. "Toward Expert Knowledge Representation for Automatic Breast Cancer Detection", in: AIMSA '08: Proceedings of the 13th international conference on Artificial Intelligence, 2008, pages 333–344.

M. Samulski, N. Karssemeijer, P. Lucas and P. Groot. "Classification of mammographic masses using support vector machines and Bayesian networks", in: Medical Imaging, volume 6514 of Proceedings of the SPIE, 2007, page 1, page 65141J.

Other publications

M. Velikova, N. Ferreira, **M. Samulski**, P. Lucas and N. Karssemeijer. "An Advanced Probabilistic Framework for Assisting Screening Mammogram Interpretation". Book Chapter. In: Computational Intelligence in Healthcare 4, volume 309, pages 371–395, Editors: I. Bichindaritz, S. Vaidya, A. Jain, and L. Jain, Springer Berlin/Heidelberg 2010. ISBN: 978-3-642-14463-9.

N. Karssemeijer, **M. Samulski**, G. den Heeten and C. Boetes. "Analysis of Observer Performance Based on Probing Patterns in an Interactive CAD System for Mammographic Mass Detection", Scientific Presentation at RSNA 2008 94th Scientific Assembly and Annual Meeting, McCormick Place, Chicago, 2008.

N. Karssemeijer, **M. Samulski**, M. Kallenberg, A. Hupse, C. Boetes and G. den Heeten. "Effectiveness of an Interactive CAD System for Mammographic Mass Detection", Scientific Presentation at RSNA 2008 94th Scientific Assembly and Annual Meeting, McCormick Place, Chicago, 2008.

M. Samulski, N. Karssemeijer, C. Boetes and G. den Heeten. "An Interactive Computer-aided Detection Workstation for Reading Mammograms". Education Exhibit at RSNA 2008 94th Scientific Assembly and Annual Meeting, McCormick Place, Chicago, 2008.

Summary

For many years, it has been recognized that even the best radiologists make errors when reading medical exams including perception failures and interpretation failures. To reduce these problems, computer aided detection and diagnosis systems have been designed to aid radiologists detecting and classifying abnormalities. The first part of this thesis concerns combining information from multiple mammographic projection views to improve detection performance of computer aided detection systems. Most computer-aided detection systems that are used in the clinic today are focussed on reducing perception errors. The research presented in the second part of this thesis investigates if presenting CAD results in a fundamentally different way to avoid interpretation errors is more effective than current computer aided detection methods that focus on preventing perceptual oversights in medical screening.

In Chapter 2, two machine learning techniques, namely support vector machines and Bayesian networks, were evaluated for characterizing masses as either benign or malignant. In addition, the effectiveness of dimension reduction (principal component analysis) and normal distribution transformation (Manly transformation) were investigated. It was found that the area under the ROC curve (A_z) of the naive Bayesian classifier increased significantly (p=0.0002) when the Manly transformation was used, from $A_z = 0.767$ to $A_z = 0.795$. The Manly transformation did not result in a significant change for support vector machines. The difference between the support vector machines and the naive Bayesian classifiers using the transformed data set was not statistically significant (p=0.78). Applying dimension reduction in the form of PCA improved the classification accuracy of both classifiers, but the difference between the two classifiers after applying PCA was not statistically significant.

In a breast screening program, it is important to combine all available information from a patient for making a referral decision. In Chapter 3 a Bayesian network framework was proposed that exploits multi-view dependencies for the analysis of screening mammograms. Instead of focussing on improving the localized detection of breast cancer, a CAD system was built that discriminates between normal and cancerous patients. It was investigated whether a reliable likelihood measure for a patient being cancerous could be obtained by combining information available as detected regions from a single-view CAD system from both mammographic views. This approach was tested with screening mammograms for 1063 patients of whom 385 had breast cancer. The results show that the multi-view modeling lead to significantly better performance in discriminating between normal and cancerous patients compared to using a single-view CAD system.

During screening, a medio-lateral oblique (MLO) and cranio caudal (CC) view are often obtained from both breast. To train CAD systems that use correspondence information, corresponding regions in those views have to be found. Therefore, in Chapter 4

a method was developed to classify region pairs into the four possible types (combinations between a TP region in both views, a combination between a TP and FP or FP and TP, and combinations between FP regions). For each combination between regions some similarity features are calculated such as the difference in distance to the nipple, grayscale correlation, histogram correlation and other features that indicate similarity. Using these features, a 4-class k-Nearest Neighbour classifier was trained, to determine the four likelihoods for each combination type. The method was tested on an annotated dataset with 412 cases. Results show that for 82.4% of the TP regions, a correct link could be established. The difference between the 4-class kNN classifier and the LDA classifier from previous research was not statistically significant. When choosing threshold such that the percentage of correct TP-TP combinations is 70%, the number of TP-FP combinations decreases significantly when using the 4-class kNN classifier (Fisher's exact test, $p \leq 0.0008$). It is expected that the decrease in TP-TP combinations will have a less negative effect on the detection performance of the two-view classifier because the regions are independently analyzed by the single-view CAD system.

Radiologists generally combine information from multiple views to detect suspicious regions in mammograms. However, most of the current CAD systems analyze each view independently. It was investigated if case-based detection performance could be improved by optimizing the learning process of the multi-view classifier. Based on the output of a correspondence classifier that classified region pairs into the four possible types, the selection of training patterns to train the multi-view CAD system was biased. In that way, the training could be focussed towards improvement of case-based detection performance. The method was tested on 454 mammograms consisting of 4 views with a malignant region visible in at least one of the views. Case-based evaluation showed a mean sensitivity improvement of 4.7% in the range of 0.01-0.5 false positives per image.

Mammographic CAD systems that are currently used in clinical practice focus only on the problem of perception errors; however, misinterpretation is a far more common cause of missing breast cancer in screening than perceptual oversights. In Chapter 6 it was investigated if a workstation that allows readers to probe image locations for the presence of CAD information while reading mammograms could improve detection performance. If a CAD finding was present on the queried location, it was displayed with the computer estimated malignancy score. The approach was evaluated using an observer study with nine readers (four screening radiologists and five non-radiologists). The participants read 120 cases of which 40 cases had a malignant mass that was missed at the original screening. The performance of the average reader significantly increased with interactive CAD at low false-positive rates from 25.1% to 34.8%, without affecting reading time. It was found that in addition to using CAD in

the traditional way to avoid perception errors, there is a large potential for using CAD as a decision aid to reduce interpretation failures.

In Chapter 7 it was investigated if the interactive computer-aided detection (CAD) system introduced in 6 increases mass detection performance in comparison to the regular CAD prompting systems currently used in clinical practice. An observer study was conducted in which six certified screening radiologists and three residents read 200 difficult cases. Results show that the reader sensitivity increased significantly ($p < 0.01$) when interactive CAD was used (58.5%) compared to both reading without CAD (51.2%) and reading with CAD prompts (51.1%). There was no significant difference found in the number of unreported abnormal cases when mammograms were read with interactive CAD compared to reading with prompting CAD or to reading without CAD.

In the last chapter it was investigated if the interactive method of presenting CAD results could also improve the usefulness of CAD in another application, namely the detection of nodules in chest radiographs. The effect of prompts and interactive use of CAD for detecting chest nodules was compared. Six readers read 247 chest radiographs that were selected from the publicly available JSRT database. The CAD results were taken from the commercially available CAD system (Riverain OnGuard™5.0). It was shown that with CAD prompting, mean sensitivity of the readers increased significantly from 35.2% to 42.8%. When using interactive CAD, the performance of the average reader increased significantly to 49.5%. This showed that CAD as a decision aid can improve readers' nodule detection performance compared to the traditional use of CAD prompts, in particular at low false positive rates.

Samenvatting

Al vele jaren wordt erkend dat zelfs de beste radiologen fouten maken bij de beoordeling van radiologische beelden, waaronder perceptie fouten en interpretatie fouten. Om deze problemen te verminderen, zijn computer-ondersteunde detectie (CAD) and diagnose systemen ontworpen om radiologen te helpen met het detecteren en classificeren van afwijkingen. Het eerste deel van dit proefschrift heeft betrekking op het combineren van informatie uit meerdere mammografische opnamerichtingen om de prestaties van een CAD systeem te verbeteren. De meeste CAD systemen die hedendaags worden gebruikt in de kliniek zijn gericht op het verminderen van perceptie fouten. Het onderzoek beschreven in het tweede deel van dit proefschrift bestudeert of de presentatie van CAD resultaten op een fundamenteel andere wijze om te voorkomen dat interpretatie fouten worden gemaakt effectiever is dan de huidige CAD systemen die zich richten op het voorkomen dat een tumor wordt overzien.

In hoofdstuk 2 werden twee machine learning technieken, support vector machines en Bayesiaanse netwerken, geëvalueerd voor het karakteriseren van tumorschaduwen als goedaardig of kwaadaardig. Daarnaast werd de effectiviteit van dimensie reductie (principal component analysis) en de normale verdeling transformatie (Manly transformatie) onderzocht. Er werd vastgesteld dat de oppervlakte onder de ROC curve (A_z) van de naïeve Bayesiaanse classifier significant toenam (p=0,0002) als de Manly transformatie werd gebruikt, van $A_z = 0,767$ tot $A_z = 0,795$. De Manly transformatie resulteerde niet in een significante verandering voor support vector machines. Het verschil tussen de support vector machines en de naïeve Bayesiaanse classifiers met behulp van de getransformeerde data set was niet statistisch significant (p = 0,78). Toepassing van dimensie reductie in de vorm van PCA verbeterde de classificatie nauwkeurigheid van beide classifiers, maar het verschil tussen de twee classifiers na deze dimensie reductie was niet statistisch significant.

In een bevolkingsonderzoek programma is het belangrijk om alle beschikbare informatie van een patiënt te combineren om tot een beslissing te komen om de patiënt te verwijzen of niet. In hoofdstuk 3 werd een Bayesiaans netwerk raamwerk voorgesteld dat afhankelijkheden tussen meerdere radiografische opnamen gebruikt voor de analyse van mammogrammen verkregen tijdens het bevolkingsonderzoek. In plaats van te focussen op het verbeteren van het exact lokaliseren van borstkanker, werd een CAD-systeem ontwikkeld dat discrimineert tussen patiënten met en zonder kanker. Er werd onderzocht of een betrouwbare maat kan worden verkregen voor de kans dat een patiënt kanker heeft door het combineren van beschikbare informatie zoals gedetecteerde regio's uit een single-view CAD-systeem van beide mammografische opnamerichtingen. Deze aanpak werd getest met screeningsmammogrammen voor 1063 patiënten, van wie 385 borstkanker had. De resultaten laten zien dat deze methode leidt tot een significant betere prestatie in het onderscheid maken tussen normale

en kanker patiënten vergeleken met het gebruiken van een single-view CAD systeem.

Tijdens de screening wordt vaak een medio-laterale oblique (MLO) opnamerichting en een cranio caudaal (CC) opnamerichting verkregen van beide borsten. Om CAD-systemen te trainen die correspondentie informatie gebruiken, moeten regio's die bijelkaar horen worden gevonden in beide opnamerichtingen. Daarom is in hoofdstuk 4 een methode ontwikkeld om regioparen in te delen in de vier mogelijke types (combinaties tussen een TP regio in beide opnamerichtingen, een combinatie tussen een TP en FP of FP en TP, en de combinaties tussen FP regios). Voor elke combinatie tussen de regio's werden gelijkheidskenmerken berekend, zoals het verschil in afstand tot de tepel, grijstinten correlatie, histogram correlatie en andere kernmerken die gelijkenis aangeven. Met behulp van deze kenmerken, werd een 4-klasse k-Dichtsbijzijnde Buren classifier getraind, om de vier waarschijnlijkheden voor elke combinatie type te bepalen. De methode werd getest op een geannoteerde dataset bestaande uit 412 patiënten. De resultaten tonen aan dat voor 82.4 % van de TP's, een juiste link kon worden vastgesteld. Het verschil tussen de vier-klasse KNN classifier en de LDA classifier uit eerder onderzoek was niet statistisch significant. Bij de keuze van de drempel zodat het percentage van de juiste TP-TP combinaties 70 % is, vermindert het aantal TP-FP combinaties significant bij het gebruik van de 4-klasse KNN classifier (Fisher's exact test, $p \leq 0,0008$). De verwachting is dat de reductie van het aantal TP-TP combinaties een minder negatief effect zal hebben op de tumordetectie prestaties van de multi-view classifier, omdat de regio's dan onafhankelijk worden geanalyseerd door het single-view CAD-systeem.

In de dagelijkse praktijk combineren radiologen informatie uit meerdere opnamerichtingen (mediolateral oblique (MLO) en cranio-caudal (CC) opnames) om verdachte tumors te detecteren in mammogrammen. Echter de meeste van de huide CAD systemen analyseren elke opnamerichting onafhankelijk. Er werd onderzocht of de patiëntgebaseerde prestaties can een CAD systeem verbeterd zou kunnen worden door het optimaliseren van het leerproces van een 'classifier', een computer-programma afkomstig uit de kunstmatige inteligentie die probeert de cognitieve vaardigheden van de mens te evenaren of overtreffen. Op basis van de informatie van een correspondentie classifier die de combinaties van verdachte regio's probeert in te delen in de vier mogelijke gevallen, wordt een bepaalde selectie van regio's met hun kenmerken aangeboden om een multi-view CAD systeem te trainen. Door een bepaalde selectie van patronen aan te bieden, zou men het leren van de classifier kunnen focussen op het verbeteren van patiënt-gebaseerde prestaties, dat wil zeggen dat het detecteren van de tumor in een van de opnamerichtingen volstaat. Deze methode werd getest op mammogrammen van 454 patiënten. Van elke patiënt zijn er 4 beelden (2 borsten in elk 2 opnamerichtingen) beschikbaar met een kwaadaardige regio zichtbaar in ten min-

ste een van de opnamerichtingen. Uit de patiënt-gebaseerde evaluatie bleek dat de gemiddelde sensitiviteit verbetert van 4,7 % in het bereik van 0,01 tot en met 0,5 false positives per beeld.

Mammografische CAD-systemen die momenteel worden gebruikt in de klinische praktijk richten zich uitsluitend op het probleem van de perceptie fouten, maar verkeerde interpretatie is een veel meer voorkomende oorzaak van het missen van borstkanker bij het screenen dan perceptuele vergissingen. In hoofdstuk 6 werd onderzocht of een werkstation die het mogelijk maakt om tijdens het lezen van de mammogrammen CAD informatie op te vragen door met de computermuis te klikken op verdachte gebieden, de tumordetectie prestaties kan verbeteren. Als er CAD informatie beschikbaar was op de opgevraagde locatie, dan werd de CAD prompt getoond als een gekleurde contour (variërend van geel tot rood) met een door de computer geschatte kans op maligniteit. De aanpak werd geëvalueerd met behulp van een waarnemer studie met negen lezers (vier screening radiologen en vijf niet-radiologen). De deelnemers lazen 120 gevallen, waarvan 40 gevallen een kwaadaardige tumorschaduw hadden die werd gemist tijdens de oorspronkelijke screening. De sensitiveit van de gemiddelde lezer nam significant toe met het gebruik van interactief CAD, bij een lage foutpositieve doorverwijzingspercentage, van 25,1 % tot 34,8 %, zonder dat dit invloed had op de leestijd. Er werd gevonden dat, naast het gebruik van CAD op de traditionele manier om perceptie fouten te vermijden, er een groot potentieel is voor het gebruik van CAD als een beslissingsondersteuning om interpretatie fouten te verminderen.

In hoofdstuk 7 werd onderzocht of het interactieve CAD systeem geintroduceerd in hoofdstuk 6 leidt tot een hogere tumorschaduw detectie performance in vergelijking met de traditionele CAD prompt systemen die momenteel in de klinische praktijk worden gebruikt. Een waarnemer onderzoek werd uitgevoerd waarin zes gecertificeerde screening radiologen en drie residents 200 moeilijke gevallen hebben gelezen. De resultaten tonen aan dat de sensitiviteit van de lezer significant toenam ($p < 0,01$) wanneer interactief CAD werd gebruikt (58,5 %) ten opzichte van zowel het lezen zonder CAD (51.2 %) en het lezen met CAD-prompts (51.1 %). Er werd geen significant verschil gevonden in het aantal niet-gemelde gevallen wanneer abnormale mammogrammen werden gelezen met interactieve CAD in vergelijking met het lezen met prompting CAD of het lezen zonder CAD.

In het laatste hoofdstuk werd onderzocht of de interactieve manier van het presenteren van CAD resultaten ook het nut van CAD zou kunnen verbeteren in een andere toepassing, namelijk bij het opsporen van knobbeltjes (nodules) in borstkas röntgenfoto's. Het effect van CAD prompts en interactief gebruik van CAD voor het opsporen van long nodules werd vergeleken. Zes lezers lazen 247 thoraxfoto's die werden geselecteerd uit de openbaar beschikbare JSRT database. De CAD-resultaten waren afkom-

stig van het commercieel beschikbare CAD-systeem (Riverain OnGuard TTra 5.0). Er werd aangetoond dat met CAD prompts, de gemiddelde sensitiviteit van de lezers significant toenam van 35,2 % tot 42,8 %. Bij het gebruik van interactief CAD, nam de prestatie van de gemiddelde lezer significant toe tot 49,5 %. Daaruit bleek dat interactief CAD als beslissingsondersteuning de nodules detectie prestaties van lezers kan verbeteren in vergelijking met het traditionele gebruik van CAD-prompts, vooral bij een laag percentage fout-positieven.

Dankwoord

Eindelijk is het zover, het boekje is af! Hoewel alleen mijn naam op de kaft staat, had ik dit proefschrift niet kunnen schrijven zonder de hulp van vele collega's, vrienden en familie. Ik heb met velen gediscussieerd, hulp gekregen bij het oplossen van lastige problemen, en morele steun gekregen op momenten dat het even tegenzat. Al deze mensen wil ik hartelijk bedanken. Graag noem ik aantal mensen bij hun naam en hoop dat ik niemand zal vergeten.

Allereerst wil ik mijn promotoren, Nico Karssemeijer en Peter Lucas, bedanken voor de mooie tijd in Nijmegen. Nico, bedankt dat jij me de kans gegeven hebt om na mijn master traject, ook mijn promotie onderzoek te kunnen doen in jouw groep. Ik heb erg veel van je geleerd, je deur stond altijd open, en zonder je raad en vele ideeën zou het me niet zijn gelukt. Peter, allereerst wil ik je bedanken dat je me hebt geïnspireerd om onderzoek te doen. Door jou ben ik in contact gekomen met Nico Karssemeijer, en ik heb met veel plezier met je samengewerkt in het B-SCREEN project.

Een goede gezellige sfeer is belangrijk in je werk, en wat dat betreft heb ik het getroffen met mijn collega's. Voor een groot deel van mijn promotie had ik het geluk de kamer te mogen delen met Pieter Vos. Pieter, naast de werk discussies, wil ik je vooral bedanken voor de lol die we hebben gehad in de barak. Michiel, Guido en Rianne, onze promotie trajecten kende vele raakvlakken en we konden daardoor veel inhoudelijk discussieren, maar vooral de informele gesprekken en gezelligheid herinner ik me nog het meest. Ik wens jullie veel succes met het afronden van jullie promotieonderzoek. Guido van Schie, bedankt dat je mij tijdens de promotie als paranimf wilt bijstaan, met alle daar bij komende zaken. Mede dankzij jou was de sfeer in de barak erg plezierig. En Rianne, samen hebben we toch mooi onderzoek gedaan dat tot een hoofdstuk in dit en jouw proefschrift heeft geleid.

Marina Velikova, ik bedank je voor de leuke onderzoekstijd die we samen hebben beleefd. De ervaring die jij al had opgedaan toen je begon als post-doc in het B-SCREEN project was voor mij erg waardevol. Ik vond het erg leuk om met jou samen te werken.

Als ik in de knel zat met wiskundige problemen kon ik altijd terecht bij Peter Snoeren. Peter, bedankt voor het inzichtelijk maken van de wiskundige aspecten en voor de nuttige adviezen die je altijd paraat had. Mede dankzij jouw hulp is er een mooi hoofdstuk over het gebruik van interactief CAD bij het beoordelen van borstkas röntgenfoto's in dit proefschrift gekomen.

Ik wil de leden van de manuscript commissie, prof. Tom Heskes, prof. Boudewijn Lelieveldt, en prof. Mathias Prokop bedanken voor hun bereidheid in de manuscript-commissie zitting te nemen en voor de tijd die zij vrij hebben willen maken voor de kritische beoordeling van dit manuscript.

Familie, schoonfamilie, vrienden en bekenden. Bedankt voor jullie interesse in mijn

werk en voor julie steun. In het bijzonder wil ik Christian Gilissen bedanken, alweer meer dan 10 jaar een heel goede vriend. Onze educationele paden verliepen redelijk parallel. Toen we beiden afstudeerden aan de hogeschool Zuyd wist ik nog niet zeker of ik verder wilde studeren aan de universiteit. Dankzij jou ben ik er toch over na gaan denken, en heb geen spijt gehad van de beslissing om verder te gaan studeren in Nijmegen. Onze afspraken zijn misschien niet meer zo regelmatig als vroeger, de kopjes koffie, huisbezoekjes en gesprekken zijn erg belangrijk voor mij geweest. Ik ben blij dat je me wil bijstaan als paranimf, en hoop dat we in contact blijven, ondanks dat ik ben geëmigreerd naar Noorwegen.

Ook mogen in deze lijst mijn ouders en zus niet ontbreken. Lieve pap en mam, jullie hebben mij in allerlei opzichten gesteund, door jullie liefde en trots, dankzij jullie ben ik zover gekomen. Monique, ik ben heel blij dat je mijn zus bent. Bedankt voor de goede gesprekken, plezier en afleiding. Mijn schoonfamilie, Kari en Håkon, bedankt voor jullie interesse en de altijd gezellige tijd in Noorwegen.

En als laatste, mijn allerlieftste Anne-Gudrun. Met je geweldige gevoel voor humor en het vertrouwen in mij lever je een onmisbare bijdrage aan mijn leven. Het laatste jaar was erg hectisch: het promotieonderzoek moest afgerond worden en ik wilde zo snel als mogelijk emigreren naar Noorwegen om eindelijk samen te zijn. Nu is het boekje klaar, en breken er voor ons wat rustigere tijden aan. Bedankt voor alle steun. Ik hou van je!

Curriculum Vitae

Maurice René Marina Samulski was born on the 26th of December 1978 in Kerkrade, the Netherlands. After completing secondary school he started in 1995 at the Arcus College in Heerlen to study automation and electronics for 4 years. After that he studied technical computer sciences at the Hogeschool Zuyd in Heerlen and obtained his BSc degree in 2003. Then, in 2003, he started with a computer science masters programme at the Radboud University Nijmegen and wrote his master thesis on the classification of breast lesions at the Radboud University Nijmegen Medical Centre in 2006.

In september 2006 he started a Ph.D. project under supervision of dr. Nico Karssemeijer at the radiology department of the Radboud University Medical Centre in Nijmegen. His research focused on interactive computer-aided detection methods in medical screening and resulted in this manuscript. In April 2011 he emigrated to Norway to live with his fiancée Anne-Gudrun Klæth Lyngsmo.