The following full text is a publisher's version.

For additional information about this publication click this link.
http://hdl.handle.net/2066/99158

Please be advised that this information was generated on 2019-08-05 and may be subject to change.
Detection of malignant masses in breast cancer screening by computer assisted decision making

Rianne Hupse
DETECTION OF MALIGNANT MASSES IN BREAST CANCER SCREENING BY COMPUTER ASSISTED DECISION MAKING

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
DONDERDAG 18 OKTOBER 2012 OM 11.00 UUR PRECIES

DOOR

ANDREA MARIANNE GERRARDA HUPSE

GEBOREN OP 3 JUNI 1981 TE GOUDA
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 260.

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”.

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

1 Introduction  
1.1 Computer-aided detection in breast cancer screening  2  
1.2 The effect of computer-aided detection  3  
1.3 New developments  9  
1.4 Outline of this thesis  10  

2 Use of normal tissue context in computer-aided detection of masses in mammograms  
2.1 Introduction  17  
2.2 Materials and methods  19  
2.3 Results  29  
2.4 Discussion and conclusion  36  

3 The effect of feature selection methods on computer-aided detection of masses in mammograms  
3.1 Introduction  43  
3.2 Methods  44  
3.3 Results  51  
3.4 Discussion and conclusion  54  

4 Using computer aided detection in mammography as a decision support  
4.1 Introduction  61  
4.2 Materials and methods  62  
4.3 Results  66  
4.4 Discussion  72  
4.5 Conclusions  76  

5 Computer aided detection of masses in mammography: interactive decision support versus prompting  
5.1 Introduction  81  
5.2 Materials and methods  82  
5.3 Results  86  
5.4 Discussion  87
6  Dual stage presentation of computer-aided detection marks in mammography to reduce both perceptual and interpretation errors 95
6.1  Introduction ......................................................... 97
6.2  Methods ............................................................ 97
6.3  Results ............................................................. 98
6.4  Discussion ........................................................ 101

7  Computer aided detection of masses in mammography: interactive decision support versus independent combination of reader and computer scores 105
7.1  Introduction ......................................................... 107
7.2  Methods ............................................................ 108
7.3  Results ............................................................. 111
7.4  Discussion and conclusion ...................................... 113

8  Standalone computer-aided detection compared to radiologists’ performance for the detection of mammographic masses 117
8.1  Introduction ......................................................... 119
8.2  Material and methods ............................................. 120
8.3  Results ............................................................. 124
8.4  Discussion ........................................................ 127

Publications .......................................................... 135

Summary ............................................................ 137

Samenvatting ........................................................ 141

Dankwoord .......................................................... 145

Curriculum Vitae .................................................... 146
Introduction
1.1 Computer-aided detection in breast cancer screening

To detect breast cancers in an early stage, in most western countries screening programs are organized. Early detection is important because it improves the chance for full recovery. In the screening programs, women in a given age group are regularly invited (every year or every two years) to obtain a mammographic screening. The large number of mammograms acquired are read by radiologists. A small fraction of these mammograms contain malignancies, which can be very subtle in an early stage. In order to avoid oversight errors, computer-aided detection (CAD) systems have been developed. These systems help to find abnormalities in the breast and therefore act as perception aid. CAD algorithms detect suspicious regions in a digitized or digital mammogram based on several image characteristics. Suspicious regions are marked by the system with prompts, which can be switched on and off by the radiologist. The radiologist can review these regions and makes a final decision whether further assessment is necessary. Mammographic signs of cancer can roughly be divided into two groups: microcalcifications and masses.

1.1.1 Microcalcifications

Microcalcifications are calcium deposits that appear as small (between 0.1 and 0.5 mm) white specks on the mammogram. They can be distributed in one or more clusters, fill a segment of the breast or are scattered over the whole breast. Most microcalcifications indicate a benign process, however, they can also be an early sign of breast cancer. Malignant microcalcifications appear typically in a cluster and have irregular, pleomorphic shapes. An example of a malignant cluster of microcalcifications is shown in figure 1.1a.

1.1.2 Masses

A mass is a space occupying lesion with a circumscribed, indistinct or spiculated margin. Spiculation is a stellate pattern of lines directed towards the mass centre and is an important sign of malignancy. An example of a spiculated mass is shown in figure 1.1b. Malignant masses are often accompanied by microcalcifications. Sharp, circumscribed borders are often an indication that the mass is benign. In practice, often the term “masses” is used for the whole group of masses, architectural distortions and asymmetric densities. An architectural distortion is an interruption of the normal ductal pattern in the breast, with no definite mass visible. This includes spiculations and focal retraction of the edge of the breast tissue. Asymmetric densities are visible as
asymmetry of tissue density of the left and right breast.

1.2 The effect of computer-aided detection

1.2.1 Performance measurements in breast cancer screening

Each decision in screening can be classified in one of four groups: true-negatives (TN), false-positives (FP), false-negatives (FN) and true-positives (TP). This classification is based on the decision outcome (recall or not) and the actual state of the case (cancer or not). The numbers of decisions in each group form together a confusion matrix as shown in table 1.1. A screening program is considered to have a good performance when the numbers for true-positives and true-negatives are large and the numbers for false-positives and false-negatives are small. To measure this performance, several measurements can be used:

- The true-positive rate or sensitivity: \( \frac{TP}{TP + FN} \)

<table>
<thead>
<tr>
<th>Actual state:</th>
<th>Decision outcome:</th>
</tr>
</thead>
<tbody>
<tr>
<td>No cancer</td>
<td>No recall - # true-negatives (TN)</td>
</tr>
<tr>
<td>Cancer</td>
<td># false-negatives (FN)</td>
</tr>
</tbody>
</table>

Figure 1.1: Two examples of mammographic signs of cancer: a cluster of microcalcifications (a) and a spiculated mass (b).
Introduction

- The false-positive rate: \( \frac{FP}{FP + TN} \)
- The specificity: \( 1 - \text{false-positive rate} \)
- The cancer detection rate: \( \frac{TP}{TP + FN + TN + FP} \)
- The recall rate: \( \frac{FP + TP}{TP + FN + TN + FP} \)
- The positive prediction value (PPV), which is the cancer detection rate divided by the recall rate: \( \frac{TP}{FP + TP} \)

Some other terms that are used in studies on screening performance are:

- The number of cancers in a screening population: \( TP + FN \)
- The prevalence of cancer in a screening population: \( \frac{TP + FN}{TP + FN + TN + FP} \)

The determination of the “actual state” is not trivial when the case is not recalled. For counting the number of false-negatives, one could take all cancers detected in the period between the current screening round and the next round (interval cancers). Another way is to check the mammograms retrospectively for abnormalities for the cases in which cancer was detected in the next screening round.

1.2.2 Studies that showed a positive effect

Several studies showed a positive effect of CAD in breast cancer screening, however, there are also a number of studies that do not show a positive effect. Therefore, the effect of CAD techniques that are currently used in screening programs is not conclusive.

Studies that measured a positive effect of CAD in breast cancer screening include studies by Freer\(^1\), Helvie\(^2\), Birdwell\(^3\), Morton\(^4\), Dean\(^5\) and Ko\(^6\). These six studies were prospective studies in which mammograms were interpreted first without CAD and then with the aid of CAD. Decisions of the radiologists were recorded before and after display of the CAD findings. The studies use a “cross-sectional design”, because all mammograms are interpreted twice at one specific point in time. Most of the six studies show a significant increase in the number of detected cancers, and a small increase in recall rate. Cancer detection rates were increased by respectively 19.5\%, 10.0\%, 7.4\%, 7.6\%, 12.9\% and 4.7\%. The absolute increases in recall rates were 1.2\%, 1.4\%, 0.8\%, 1.0\%, 1.6\% 2.0\%. Further, Dean and Ilvento\(^5\) found that the additional detected cancers were significantly smaller. Freer and Ulissey\(^1\) found an increase in the proportion of early stage malignancies.
1.2 The effect of computer-aided detection

Another study that measured a positive effect of CAD is the CADET II study, performed by Gilbert et al.\textsuperscript{7}. In this large multi center prospective study more than 31,000 mammograms were read in two modes: double reading without CAD and single reading with CAD. No sequential recording of reading without and with CAD was used. Instead, each reader read a series of mammograms either with CAD or without CAD. For double reading, mammograms with discordant results were arbitrated by a third reader or another pair of readers. A small amount of mammograms (2334) were read in either one of the modes and were not included in the analysis. The majority of mammograms (more than 28,000) were independently read in both modes by different readers. For these mammograms, single reading with CAD was compared to double reading. No significant difference in cancer detection rate was found for this comparison. The recall rate for single reading with CAD was slightly higher (0.5%, significant) than for double reading. Results indicate that single reading with CAD could be an alternative to double reading.

The studies performed by Cupples et al.\textsuperscript{8} and Gromet\textsuperscript{9} used a “longitudinal design“. They analyzed the number of cancer detections and recalls in a screening program before and after the introduction of CAD. An increase in cancer detection rate of respectively 16.1% and 2.0% and an absolute increase in recall rate of 0.6% and 0.4% was found. Although the increase in detection rate in the study by Gromet was not significant, a significant increase in sensitivity was found (81.4% to 90.4%). Further, in the study by Cupples, the mean age at screening detection was 5.3 years younger when CAD was used and in multi-variable analysis there was a strong association between early stage cancer and detection by CAD.

1.2.3 Studies that did not show a positive effect

Studies that did not show a positive effect include studies by Gur\textsuperscript{10}, Georgian-Smith\textsuperscript{11} and Fenton\textsuperscript{12,13}. In the study by Georgian-Smith et al.\textsuperscript{11} sequential reading without and with CAD was used. No increase in detection rate was found while recall rate increased by 0.5%. The studies by Gur\textsuperscript{10} and Fenton\textsuperscript{12,13} were based on a longitudinal design. Gur found an increase in detection rate of 1.7% (not significant) and no effect on the recall rate.

Recently, studies by Fenton et al.\textsuperscript{12,13} led to a huge debate on the overall efficacy of CAD systems. In the first study in 2007, records from more than 220,000 women were analyzed who received mammograms at screening facilities in the United States. The second study in 2011 included data from more than 680,000 women. Both studies showed an increased recall rate when CAD was introduced (3.1% and 0.5%). However, no significant increase in cancer detection rate was found (an increase of 1.2% and a
decrease of 11.1%), nor in sensitivity (3.6% and 1.4%). The positive prediction value
(PPV) is a measure for the ratio between the number of detected cancers and recall
rate and measures the probability that a recalled woman has a cancer. In both Fenton
studies a decrease in PPV was observed after CAD implementation. In the 2011 study,
CAD was not associated with favorable stage, size or lymph node status of invasive
breast cancer. Further, the small increase in sensitivity was attributed to an increased
detection of ductal carcinoma in situ (DCIS). Low grade DCIS is considered a relatively
indolent type of cancer.

1.2.4 Interpretation of contradicting study results

The described studies on the effect of CAD used multiple study designs (longitudinal
and cross-sectional). Further, different measurements like cancer detection rate, recall
rate, sensitivity, positive prediction value and the stage of detected cancers were used.

The longitudinal study design

The main advantage of a longitudinal study design is that a large dataset can easily
be collected retrospectively. This is important, because prevalence of breast cancer is
low (approximately 4 in a thousand screened women have breast cancer) and there-
fore statistically significant results can only be obtained with a relatively large dataset.
However, there are several important disadvantages when using a longitudinal study
design. First, data from different time periods is compared (before and after introd-
uction of CAD in screening). Therefore, study and control groups should be matched for
patient age, breast density, proportion of incident screening rounds and reader vari-
ability should be included in the analysis. Learning curve effects can also influence the
results because a radiologist can perform better over time when getting more experi-
ce with CAD. Both variability of radiologists and learning curve effects were ignored
in the studies by Fenton12,13.

A second disadvantage of the longitudinal study design is that cancer detection
rate is not a useful measurement when comparing data acquired before and after the
introduction of CAD in screening. The reason for this is as follows. The number of de-
tected cancers in a given time period depends not only on radiologist sensitivity (with
or without CAD), but also on the number of detectable cancers presented at screening
in the given time period. When radiologist sensitivity increases due to the use of CAD,
the cancer detection rate will increase in the year of introduction. However, due to the
larger number of detected cancers, the number of detectable cancers presented at the
next screening round will be lower than before the introduction of CAD. Therefore, in
1.2 The effect of computer-aided detection

the following time period the cancer detection rate will decrease again. The overall effect will be that cancers are detected in an earlier screening round. This effect was simulated by Nishikawa\textsuperscript{14} in 2007. He also showed that there is a large variability in the number of detectable cancers that are presented in screening each year. This variation is caused by the fact that the growth rate is not the same for each cancer. Another effect of introducing an effective CAD system is that the number of cancers detected in between two screening rounds (interval cancers) will decrease. Therefore, in screening, a small increase in cancer detection rate is expected. However, it is hard to detect this increase because it is relatively small compared to the variability in cancer detection rate from year to year.

The cross-sectional study design

When using the cross-sectional study design, cancer detection rate is indeed an useful measurement. This is because the interpretation of mammograms in each of the two modes (with and without CAD) is done at the same point in time. Therefore, the prevalence in the study population is constant and any difference in the number of detected cancers is directly related to a difference in reader sensitivity. Differences in detection rate can not be the effect of differences in study population because the study groups contain the same patients. When a sequential reading approach is used, results for reading with and without CAD can be compared for the same radiologists and at the same time which makes the influence of the intra-reader variance and the inter-reader variance minimal. Therefore, this type of study has the highest statistical power for showing a benefit of CAD (the statistical power is the probability that the study shows an effect of CAD while there is an effect of CAD). When focusing on the mentioned studies with cross-sectional design and sequential reading, the average increase in detection rate was 9\% (0\% to 19.5\%). This indicates a positive effect of CAD. Most of these studies also show an increase in recall rate and therefore more false-positive recalls. However, on average, the PPV for unaided reading (5.1\%) was similar to the PPV for reading with CAD (4.9\%).

When interpreting these results, one should take into account there is also a disadvantage of the study design with sequentially reading. Mammograms are read sequentially without and with CAD. Only the final decision (after display of the CAD prompts) will affect patient care. Therefore, it is not certain if decisions for the unaided mode reflect decisions that would have been made in an unaided screening setting. It might be that radiologists rely on the CAD prompts and are less vigilant in the unaided mode. This will introduce a positive bias for the effect of CAD. On the other hand, radiologists might be more alert when they know they are part of a study. This
Introduction

will introduce a negative bias for the effect of CAD. In the CADET II study by Gilbert et al.\textsuperscript{7}, any potential positive bias was minimized because each reader read each mammogram either with or without CAD. Further, a relatively small number of mammograms (not included in analysis) was only read in one of the two modes. Therefore, readers did not know whether the mammogram would also be read in the other mode.

**Sensitivity as performance measure**

In order to overcome the limitations of using cancer detection rate in a longitudinal study, one might choose to use sensitivity as performance measure. Sensitivity is defined as the fraction of cancers in the screening population that are detected. Although sensitivity is not affected by differences in prevalence in the screening groups, it is still unreliable to compare measurements from different periods in time. The reason is that it is hard to define the total number of detectable cancers in the screening population, because for this one has to know how many cases were missed in screening. In order to obtain this number, often the women are counted who received a diagnosis of breast cancer in the period between the current screening round and the next round (interval cancers). However, not all missed cancers are detected as interval cancers. In fact, many cancers do not grow that fast and are detected in the next screening round. The screening interval in the United States, where most studies were done, is only one year. If the use of CAD yields additional detected cancers that otherwise would not have been detected until the next screening round, these cancers should be counted as missed cancers in the unaided mode for a fair comparison. This approach will lower the sensitivity for that mode. In the Fenton studies\textsuperscript{12,13}, only interval cancers were considered, which introduced a positive bias for estimating the sensitivity in the unaided mode.

**Size and stage as performance measure**

Another way to examine the effect of CAD is to measure cancer size or stage at the moment of detection. These measures seem to be more relevant than detection rate and sensitivity because they are directly associated with earlier detection. Unfortunately, relatively few studies mentioned cancer size and stage and the studies that did had a variety of conclusions. In the (cross-sectional) studies by Freer\textsuperscript{1} and Dean\textsuperscript{5} it was found that the proportion of early staged malignancies increased. In the (longitudinal) study by Cupples\textsuperscript{8}, the mean age at screening detection was 5.3 years younger when CAD was used and in multi-variable analysis there was a strong association between early stage cancer and detection by CAD. However, in the study by Fenton in 2011\textsuperscript{13}, CAD was not associated with favorable stage, size or lymph node status of invasive
1.3 New developments

1.3.1 Digital mammography

The algorithms of most CAD systems are based on screen-film mammography (SFM). In SFM, radiation is absorbed by a scintillator with sends the signal as visible light to a film. The relation between optical film density and exposure values is non-linear and depends on the type of film used. In the last few years, full-field digital mammography (FFDM) systems have been developed and are increasingly used in clinical practice. In FFDM, a digital detector collects exposure values which yields a linear relation between exposure and pixel values. The main advantage of FFDM is that acquisition, image processing and display are separated. Therefore, each stage of image formation can be optimized individually and transmission, retrieval and storage of images are improved. For example, image processing techniques can be used to improve local contrast and display of the images can interactively be manipulated. Further, in digital mammography a lower average dose of radiation is needed since a higher signal-to-noise ratio is provided. Computer-aided detections algorithms can easily be incorporated because the images are already digital and are read on a computer monitor.

CAD systems trained on SFM images can not be directly used to detect breast cancer in FFDM images. The reason is that the different relation between pixel and exposure values yields different image characteristics. One could choose to retrain the CAD system using FFDM images, however, a high number of digital mammograms should be available. In practice, until a large database of FFDM images is available, a good solution is to convert FFDM images into SFM-like representations before applying the existing CAD algorithms\(^\text{15}\). The CAD findings can then be projected into the original FFDM image for display.

1.3.2 The use of multiple views by CAD

In recent years, several groups improved CAD performance for the detection of masses by the use of multiple views. A mammogram contains a maximum of four views: the medial lateral oblique (MLO) projection and the cranio caudal (CC) projection for the left and right breast. Information from the left and the right breast can be combined to detect asymmetry, which is an indicator for malignancy\(^\text{16-18}\). Information from the
MLO and CC view can be combined to see whether a space occupying lesion is detected in both views of the breast\textsuperscript{19–24}. If this is the case, the probability that the lesion is malignant is higher than when the lesion is only detected in a single view. Further, when reading mammograms, radiologists often compare the current mammogram to the mammogram obtained in the previous screening round. In this way information is acquired about the growth of a lesion. Temporal information can also be used in CAD algorithms to improve mass detection performance\textsuperscript{25,26}.

1.3.3 Computer-aided detection as decision aid

The performance of current CAD systems is relatively high for the detection of microcalcifications, which is appreciated by most radiologists. However, there is less agreement on the benefit of using CAD for the detection of masses and architectural distortions. In addition to study design limitations, disappointing results of CAD may be due to the way CAD marks are presented to the radiologists. In current CAD systems, prompts are displayed in order to avoid perceptual oversight errors. It has been shown that masses are often missed due to incorrect interpretation\textsuperscript{27,28} and that reader performance can be improved by retrospectively combining reader scores with the presence and probability of CAD mass markers\textsuperscript{29,30}. This indicates that CAD might have more effect when used as interpretation aid. Part of this thesis will describe a new approach of presenting CAD results, in which CAD marks are only displayed on demand for queried regions, together with a suspiciousness score. In this way the interactive system helps radiologists to interpret suspicious regions, instead of helping them with their initial detection.

1.4 Outline of this thesis

The objective of the work described in this thesis is to improve current CAD methods for the detection of malignant masses in screening. This is done by improving the detection algorithms itself, as well as the way of displaying CAD results to radiologists.

Chapter 2 describes new features for mass detection that make use of normal tissue context. These features are based on the suspiciousness scores assigned to areas in the mammogram that are assumed to depict normal tissue. When these scores are relatively large compared to the scores at a candidate mass location, it is more likely that the candidate mass is a false positive.

Chapter 3 addresses the problem of selecting a useful subset of features from a large number of available region descriptions. Reducing the number of features can improve
classification performance by reducing the risk of overfitting a classifier. Feature selection depends on the choice of a performance measure used as optimization criterion. We compare an optimization criterion that reflects the use of CAD in clinical practice to using a standard approach, and investigate the effect of feature selection compared to using all available features.

Chapter 4 and 5 present a new way of displaying CAD results to the radiologist. Current CAD systems make use of prompts that are intended to avoid perceptual oversight errors. However, in practice, many masses are not missed by perceptual oversight but due to incorrect interpretation. We developed a CAD system in which CAD marks and their associated suspiciousness scores remain hidden unless their location is queried by the radiologist. Chapter 4 investigates the effect of this interactive display and in chapter 5 the interactive display is compared to conventional prompting.

In chapter 6 a dual stage presentation of CAD marks is described. In the first stage, mammograms are read with the use of interactive CAD. In the second stage, non-referred mammograms in which one or more highly suspicious CAD regions occur that were not queried, are presented again to the reader, with the non-queried regions displayed as prompts. By doing this, the reader can still change his or her decision in case regions marked by CAD were overlooked. In this way, we investigate the combination of traditional and interactive CAD to reduce both perceptual and interpretation errors.

We expect that readers can benefit from CAD by incorporating the presence and scores of CAD marks in their own interpretation in an intelligent manner. In chapter 7 the interactive use of CAD is compared to independent combination of CAD and reader scores. The objective is to see whether readers obtain higher performance with interactive use of CAD results than with independent combination.

Chapter 8 compares the performance of a standalone CAD system to that of radiologists. Performances are compared at a high specificity comparable to the level used in screening practice.

Bibliography


Introduction


Introduction


Use of normal tissue context in computer-aided detection of masses in mammograms

R. Hupse and N. Karssemeijer

Original title: Use of normal tissue context in computer-aided detection of masses in mammograms

Published in: IEEE Transactions on Medical Imaging 2009;28:2033-2041
Abstract

When reading mammograms, radiologists do not only look at local properties of suspicious regions but also take into account more general contextual information. This suggests that context may be used to improve the performance of computer aided detection (CAD) of malignant masses in mammograms. In this study we developed a set of context features that represent suspiciousness of normal tissue in the same case. For each candidate mass region, three normal reference areas were defined in the image at hand. Corresponding areas were also defined in the contralateral image and in different projections. Evaluation of the context features was done using 10-fold cross validation and case based bootstrapping. Free response receiver operating characteristic (FROC) curves were computed for feature sets including context features and a feature set without context. Results show that the mean sensitivity in the interval of 0.05-0.5 false positives/image increased more than 6% when context features were added. This increase was significant (p<0.0001). Context computed using multiple views yielded a better performance than using a single view (mean sensitivity increase of 2.9%, p<0.0001). Besides the importance of using multiple views, results show that best CAD performance was obtained when multiple context features were combined that are based on different reference areas in the mammogram.
2.1 Introduction

Computer aided detection (CAD) systems are being developed to help radiologists detect abnormalities during breast cancer screening. In general two different CAD systems are used, one for detecting microcalcifications and one for detecting malignant masses in mammograms. This study concerns the development of a CAD system to detect malignant masses. We use the term ‘mass’ for the group of malignant masses, architectural distortions and focal asymmetries.

A CAD system for detecting malignant masses usually consists of multiple stages. In the initial stage a set of candidate mass locations is detected, based on image features extracted locally. Mass likelihood scores are assigned to each location in the image by a classifier that is trained with known examples. Although mass likelihood scores give an indication about the likelihood that a mass is present, the term likelihood is here loosely used. In pure statistical sense the scores should be normalized using conditional probabilities. For locations that have been assigned a high likelihood score, a candidate mass region is segmented. By adjusting the number of candidate locations, the sensitivity of the initial stage can be varied. It is common to set this sensitivity relatively high, close to 100%. However, due to misinterpretation of normal glandular tissue the number of false positive detections per image is also relatively high. In the following stages the candidate regions are classified into normal or malignant tissue by computing a malignancy score for each segmented region. This malignancy score is based on a large set of features.

Many features have been described that can be used to separate false positive detections (normal tissue) from true positive detections (malignant masses). These features include contrast, size and linear texture. Most of these features are based on information extracted from the segmented region and its local surroundings. However, it is known that radiologists take into account the whole image and other views of the same case. The benefit of doing this has been verified experimentally by van Engeland, who found that observers performed better in discriminating malignant masses from normal tissue when the whole image was shown instead of only the region of interest. To improve the false positive reduction in the last stage of our CAD system, we investigated the use of several groups of contextual features. One group of context features we described was related to the mass likelihood scores computed during the initial detection stage of the CAD system. In this research we investigate the relevance of different types of these context features more elaborately.

We propose that likelihood information extracted from other areas than the candidate region might be useful for interpretation of the candidate region. In Figure 2.1 two CAD regions are shown that have been assigned the same likelihood score during
the initial detection stage. The region depicted in the left image is a true positive and
the region in the right image is a false positive. Although the likelihood scores are
equal for these two candidate regions, the context in the likelihood images is very dif-
fferent for both images. Radiologists would consider the region in the left image more
suspicousness than the region in the right image because this region is more isolated
compared to the region on the right.

Figure 2.1: Upper row: two examples of original images. In the left image a malignant
mass is detected, indicated with an arrow. The right image is normal. In this image a
false positive detection is shown. The two detected regions have been assigned the same
likelihood score during the initial detection stage. Bottom row: the likelihood images
for both mammograms, including the segmentation of the regions. Although both CAD
regions have been assigned same likelihood score, the context in the likelihood images
is quite different.

The benefit of using likelihood scores for computing contextual information is that
these scores are obtained during the initial detection stage and are available for each lo-
cation in the image. Therefore, no additional segmentation is necessary other than seg-
mentation of the candidate mass. Previous research incorporated information based on regions segmented in a region of interest in the contralateral view\textsuperscript{6,7}. Because there is not always a relevant region present that can be adequately segmented, contextual features can not always be computed in this way. Normal tissue context based on likelihood scores that are obtained during the initial detection stage can always be computed for each desired area of interest in the case.

2.2 Materials and methods

2.2.1 The CAD system

Our CAD system consists of a preprocessing stage, an initial detection stage and an interpretation stage in which the number of false positive detected regions is reduced. A schematic overview of these stages is given in Figure 2.2. Preprocessing is done in three steps: segmentation into three areas (breast region, background and pectoral muscle), pectoral equalization and peripheral enhancement\textsuperscript{8}.

Candidate mass regions are detected by performing an initial detection as described by Karssemeijer\textsuperscript{9,10}. In this stage 5 features are computed for each location on a regular grid in the image. These features are two gradient concentration measures, two spiculation measures and one measure indicating the scale at which most spiculation is present. The features are fed into an ensemble of 5 neural networks that are each randomly initialized and trained on a small data set. For each location at the grid a likelihood score is computed by averaging the 5 network outputs. Together these likelihood scores form a likelihood image. This likelihood image is smoothed and each local maximum in the likelihood image that exceeds a threshold is selected as a candidate mass location. For each local maximum in the likelihood image a candidate mass region is segmented using a segmentation method based on dynamic programming\textsuperscript{11}.

In the interpretation stage the candidate masses are classified into normal or malignant tissue by computing a set of features for each segmented region. These features are the input for a second neural network classifier. By applying a threshold on the output of this classifier, the number of false positive detections is reduced without a large reduction in sensitivity. The features used in this stage measure region contrast, location, linear texture, density, region size and compactness. In addition to these features, the 5 gradient concentration and spiculation features and the mass likelihood score computed in the initial detection stage are also used by the second stage classifier.
Figure 2.2: Schematic overview of the CAD system consisting of three stages: a preprocessing stage, an initial detection stage and an interpretation stage.
2.2.2 Normal tissue context

Many false positives arise in normal dense tissue in the breast. When the likelihood scores assigned to normal tissue are relatively large compared to the candidate mass score, it is more likely that the candidate mass is a false positive. Therefore, we believe that CAD performance can be improved by including normal tissue context in the assignment of a malignancy score to the candidate region. We define normal tissue context as the mass likelihood scores in the mammogram outside the projection of the candidate mass.

Context features can be computed based on likelihood scores in several image areas:

1. **Context measure for immediate surroundings of the candidate mass region:**
   
The fact that false positive detections are often part of a larger area of dense tissue can be taken into account by providing the classifier with a feature measuring suspiciousness in the surrounding tissue of a candidate region. We define the feature ‘L-loc’ as the \( n_1 \)-th percentile of likelihood scores computed for locations within a certain distance to the segmentation border. This distance is chosen to be two times the radius of the segmented region \( r_{\text{reg}} \), which is computed by
   
   \[
   r_{\text{reg}} = \sqrt{\frac{a_{\text{reg}}}{\pi}}.
   \]  
   
   In this equation, \( a_{\text{reg}} \) is the area of the segmented region.

2. **Context measure for the whole breast:**
   
   In most of the abnormal images there are only one or two abnormal masses present. This means that if there are many candidate regions detected in an image, it is expected that most of these candidate regions are false positives. Therefore, a feature measuring the overall likelihood in an image might be used for classification of a candidate region. We define the feature ‘L-ima’ as the \( n_2 \)-th percentile of likelihood scores measured in the whole breast and pectoral muscle visible in the image.

3. **Context measure for a specific band:**
   
   There are particular areas of the breast that require special attention during screening, namely the area parallel to the edge of the pectoral muscle (“milky way”) in medial lateral oblique projection and the retroglandular space (“no mans land”) in cranio caudal projection\(^{12}\). Because these areas do not contain much glandular tissue, densities found at these areas are more suspicious. The shape of these
areas can be approximated with a band of pixels in the image with similar distance to the nipple. We define the feature 'L-b' as the $n_{3}$-th percentile of likelihood scores measured in this band. The width of the band $w$ is chosen proportional to the breast size. The relation between breast size and width of the band is chosen in a way that the width of the band is a fifth of the effective width of the breast:

$$w = 0.2 \sqrt{\frac{2A}{\pi}}$$

with $A$ the area of the breast in the image (excluding the pectoral muscle). When this band has a large distance to the nipple, it is expected that not much glandular tissue is present. Therefore, the L-b feature is expected to be more useful when the distance to the nipple is also known. To take this into account we used an estimate of this distance as a feature in each feature set validation (described in section 2.2.5).

When computing the three context features, the likelihood scores assigned to locations inside the candidate region and to locations very close to the segmentation border (2.4 mm and closer) were not used.

In general, small likelihood scores are assigned to the majority of locations in an image while large likelihood scores are assigned to very few locations. We computed $n$-th percentiles for the likelihood scores because this measure is less affected by the large amount of small scores than a measure such as the mean or median.

### 2.2.3 Incorporating information from other views

Besides extracting contextual information from the image in which the candidate region is located, also other views of the same case can be used. One case contains a maximum of four views: the medial lateral oblique (MLO) projection and the cranio caudal (CC) projection for the left and right breast. In our definition normal tissue context is based on mass likelihood scores computed for locations in the mammogram other than the projection of the candidate mass. If there is another projection available for the ipsilateral breast, sometimes no information is available where the candidate mass is located in this view. Therefore, we do not use this other ipsilateral projection for computation of context features.

In this paper we investigate inclusion of normal tissue in the mammogram at hand as a reference. Therefore, we are only interested in context outside the projections of the candidate mass. Combining features from multiple projections of a lesion in different views is not the topic of this paper but has been addressed in previous research.\(^3,13–16\)

For each candidate mass, we used three different views to compute context features: the (ipsilateral) view in which the lesion is detected (IL), the view of the contralateral
2.2 Materials and methods

breast in the same projection (CL) and the view of the contralateral breast in the other projection (CLO). When multiple views were used for computing a context feature, one cumulative distribution function was computed using all views and the $n_i$-th percentile was taken from this group of scores ($i=1,2,3$ for respectively L-loc, L-ima and L-b).

In the first row of Figure 2.3 the two views are shown that were used for computing the L-loc variants. These are the ipsilateral view (IL) and the contralateral breast in the same projection (CL). The location of the circular area in the contralateral view was determined by relative coordinates computed for the location of the candidate region in the ipsilateral view. This was done by setting up an internal coordinate frame in which the y-axis is taken in parallel to the pectoral boundary (MLO view) or to the chest wall boundary (CC view). The origin of this coordinate frame was the point that had the largest distance to the skin line. Relative coordinates of region location were obtained by dividing coordinates in this internal frame by the distance from the origin to the skin. In the ipsilateral view the likelihood scores are used with a distance to the segmentation border smaller than two times the radius of the candidate region. Likelihood scores for locations inside and near the segmentation border were not used. To obtain an area in the contralateral view with approximately the same size, we used all locations inside a circle with a radius that was three times the radius of the candidate region (one radius for obtaining an area similar to the candidate region and two times the radius for obtaining the local surroundings). Information obtained from the other projection of the contralateral breast (CLO) was not used for this feature, because a suitable location in this view can not be derived from the coordinates of the region in the ipsilateral projection.

The second and third row in Figure 2.3 show the three views that were used for computing the L-ima and L-b features. For computing L-ima the mass likelihood scores obtained for the whole breast and pectoral muscle in the other views were used. For computing L-b, a band was constructed containing locations with a similar distance to the nipple in all views. The width of each bands was defined by equation 2.2 in which $A$ is the area of the breast for the appropriate view. In the ipsilateral view locations inside and near the segmentation border were not used.

For all three context features (L-loc, L-ima, L-b) we computed three variants: one using information from the image in which the candidate region was found (IL-loc, IL-ima and IL-b), one using information from the contralateral view of the same projection (CL-loc, CL-ima and CL-b) and one using all views available (AV-loc, AV-ima and AV-b). As explained, for AV-loc only the ipsilateral and contralateral view of the same projection were used, the other projection of the contralateral breast (CLO) was not used. By using multiple views, we extended the described set of context features from
Figure 2.3: First row: the surrounding tissue used for computing the L-loc feature variants for the normal case depicted in Figure 2.1. In the ipsilateral view (IL) all locations are used with a distance to the segmentation border smaller than two times the radius of the region. In the contralateral view a circular area is used with a radius that is three times the radius of the candidate region and a location defined using internal coordinate frames. Second and third row: the ipsilateral (IL), contralateral (CL) and other contralateral (CLO) view used for computing the L-ima and L-b variants. The bands used for computing the L-b variants are shown in each view. For all context features, the likelihood scores assigned to locations inside the candidate region and to locations very close to the segmentation border (2.4 mm and closer) were not used.
3 to 9 features. These 9 variants are listed in table 2.1.

Table 2.1: Variants of context features

<table>
<thead>
<tr>
<th>feature</th>
<th>ipsilateral</th>
<th>contralateral</th>
<th>all views</th>
</tr>
</thead>
<tbody>
<tr>
<td>L-loc</td>
<td>IL-loc</td>
<td>CL-loc</td>
<td>AV-loc</td>
</tr>
<tr>
<td>L-ima</td>
<td>IL-ima</td>
<td>CL-ima</td>
<td>AV-ima</td>
</tr>
<tr>
<td>L-b</td>
<td>IL-b</td>
<td>CL-b</td>
<td>AV-b</td>
</tr>
</tbody>
</table>

In the Netherlands, two mammogram views (medial lateral oblique and cranio caudal) are obtained at the initial screening, and only one view (medial lateral oblique) at the subsequent screenings, unless there is an indication that obtaining the second view would be beneficial. Therefore, there are not always 4 views present in a case. When there are only 2 views present, the other projection of the contralateral breast (CLO) is not used for computing the features in the last column of table 2.1.

2.2.4 Parameters

For computing the context features we used three parameters: \( n_1 \), \( n_2 \) and \( n_3 \) specifying the percentile of likelihood scores computed for respectively the L-loc, L-ima and L-b variants. To obtain parameter values, we computed cumulative distribution functions using the likelihood scores obtained from the local surrounding, the whole breast or the band in the ipsilateral view for 100 normal candidate regions and 100 abnormal candidate regions. The images used for computing these functions were from the data set that was used to train the neural networks for the initial detection and were not used for validation. A total of 128 images from 60 cases were used to extract the 200 regions. For each of the three feature types, we averaged the cumulative distribution functions for the 100 normal and for the 100 abnormal regions.

In Figure 2.4 the averaged cumulative distribution functions are shown. In general, for true positive candidate regions mass likelihood scores in the context were smaller than for false positive candidate regions. For each of the three context features we computed a difference function that represented the difference between the likelihood scores of the two functions for each cumulative fraction. For each of the three features we used the cumulative fraction that yielded maximal difference. This yielded the values 0.92, 0.98 and 0.97 for respectively \( n_1 \), \( n_2 \) and \( n_3 \). For the feature variants based on the contralateral view and the variants based on all views, the same values for \( n_1 \), \( n_2 \) and \( n_3 \) were used as computed using the ipsilateral view.
Figure 2.4: Average cumulative distribution functions for likelihood scores present in the surrounding tissue at a distance of 2R from the segmentation border (upper), in the whole breast (center) and in a band containing locations with similar distance to the nipple (bottom). Functions are shown for normal and abnormal regions. The difference between the two cumulative distribution functions is also shown.
2.2.5 Feature sets

We systematically compared performances obtained using several feature sets. These feature sets were created by adding one or more context features to a standard set of features. The standard feature set consisted of the 5 gradient concentration and spiculation features computed during the initial detection stage, the likelihood score for the local maximum computed during the initial detection stage, 6 contrast measures, the size of the region, circularity, 2 line concentration measures as described in\cite{17} and the estimated distance to the nipple, normalized by the breast size.

Table 2.2: Evaluated feature sets. Each set consists of the 17 standard features and the listed context features

<table>
<thead>
<tr>
<th>feature set</th>
<th>context features</th>
</tr>
</thead>
<tbody>
<tr>
<td>$F_0$</td>
<td>-</td>
</tr>
<tr>
<td>$F_{IL}$</td>
<td>IL-loc, IL-ima, IL-b</td>
</tr>
<tr>
<td>$F_{CL}$</td>
<td>CL-loc, CL-ima, CL-b</td>
</tr>
<tr>
<td>$F_{AV}$</td>
<td>AV-loc, AV-ima, AV-b</td>
</tr>
<tr>
<td>$F_{loc}$</td>
<td>AV-loc</td>
</tr>
<tr>
<td>$F_{ima}$</td>
<td>AV-ima</td>
</tr>
<tr>
<td>$F_b$</td>
<td>AV-b</td>
</tr>
<tr>
<td>$F_{loc,ima}$</td>
<td>AV-loc, AV-ima</td>
</tr>
<tr>
<td>$F_{loc,b}$</td>
<td>AV-loc, AV-b</td>
</tr>
<tr>
<td>$F_{ima,b}$</td>
<td>AV-ima, AV-b</td>
</tr>
</tbody>
</table>

The validated feature sets are described in table 2.2. The first set ($F_0$) was the standard feature set with no context features added. The next three sets were created by adding the three context features computed using respectively the ipsilateral view ($F_{IL}$), the contralateral view ($F_{CL}$) or all views ($F_{AV}$). Feature sets $F_{loc}$, $F_{ima}$ and $F_b$ were created by adding one context feature computed using all views. Feature sets $F_{loc,ima}$, $F_{SUR_b}$ and $F_{ima,b}$ were created by adding two context features computed using all views.
2.2.6 Performance validation

To obtain the performance for each feature set we performed a 10-fold cross validation using a large database. This database consisted of 3262 normal mammograms (9688 images) and 636 abnormal mammograms (2180 images). Approximately 48% of the normal cases (1582) and 71% of the abnormal cases (454) consisted of 4 views. The other mammograms consisted of two views. Approximately 33% of the abnormal cases were mammograms taken during a screening round prior to detection of the cancer by a radiologist. In each abnormal mammogram at least one biopsy proven malignant mass was visible. All images were digitized at a pixel resolution of 50\(\mu m\) and were sampled down to a resolution of 200\(\mu m\). The gray value depth was 12 bits. The malignant masses were annotated by an experienced radiologist. These annotations were used as ground truth for validation. We performed the initial detection as described in paragraph 2.2.1 on all images. The training of the initial detection network was done using a small separate data set (302 images).

In each cross validation step a neural network classifier was trained on the candidate regions in 90% of the cases and tested on the candidate regions in the other 10%. When splitting the data into a training and test set, the images belonging to the same case were assigned to the same set. The classifier consisted of 5 neural networks that were trained with a different random initialization. Each network consisted of an input layer, a hidden layer of 12 nodes and an output layer of 1 node and was trained using the back-propagation algorithm. The number of training cycles was determined by computing a learning curve on a second separate data set (224 images). The training was stopped when the performance of the network on the separate set reached a maximum. After training the networks, for each region in the test set a malignancy score was computed by averaging the 5 outputs.

After the 10-fold cross validation the malignancy scores for all regions were pooled together and a free response receiver operating characteristic (FROC) curve was computed by plotting true positive fraction against false positive (FP) rate for a series of thresholds on the malignancy score. A true positive was defined as an abnormal case that was detected by the CAD system. The criterion for detection was that the center of mass of a CAD region was inside the annotated malignant region and the malignancy score exceeded the threshold. When multiple malignant masses were present in one case, or when the same malignant mass was annotated in two views, the case was considered a true positive if at least one CAD region was in one of the annotated regions. The false positive rate was defined as the average number of CAD regions detected per image from a normal case. To obtain a single performance measure for each feature set we computed the mean sensitivity in a range of false positive levels on a logarithmic
2.3 Results

scale:

\[ S = \frac{1}{\ln 10} \int_{0.05}^{0.5} \frac{s(f)}{f} \, df, \]  

(2.3)

with \( f \) the number of false positives in normal images and \( s(f) \) the lesion sensitivity. This measure is proportional to the partial area under the FROC curve plotted on a logarithmic scale. We chose the false positive range from 0.05 to 0.5 false positives/image as this corresponds best to the setting of current CAD systems in screening practice. Use of a logarithmic scale avoids that the measure is dominated by operating points at high false positive rates.

2.2.7 Statistical analysis of performance differences

We determined significance of the obtained performance differences between the feature sets using the bootstrap method\textsuperscript{18,19}. Cases were sampled with replacement from the complete cross validation data set 5000 times. Each new set of sampled cases contained the same number of cases as the original set. For each resampling two FROC curves were constructed using the malignancy scores obtained for the two feature sets to be compared. Subsequently the difference in mean sensitivity \( S \) was computed. Resampling 5000 times resulted in 5000 values for \( \Delta S \). P-values were defined as the fraction of \( \Delta S \) values that were negative or zero.

We performed a total of 22 comparisons. First of all, we compared the performances obtained using the following feature sets: the standard feature set (\( F_0 \)), the set with context features computed using the ipsilateral view (\( F_{IL} \)), the set with context features computed using the contralateral view (\( F_{CL} \)) and the set with context features computed using all views (\( F_{AV} \)). This yielded 6 comparisons. Secondly, we compared performances between all combinations in which 1 or 2 context features were present in the feature set (6 feature sets, 15 comparisons). Finally, the feature set in which 1 or 2 context features were present that yielded best performance was compared to feature set \( F_{AV} \) (in which three context features were present). Because of this large amount of comparisons, we applied the Bonferroni correction to the statistical significance level. Performance differences were considered significant if \( p < 0.0023 \) (0.05/22).

2.3 Results

Results of the comparisons between feature sets are listed in tables 2.3 and 2.4. In both tables, the second column shows the mean sensitivity measure \( S \) obtained for the given feature set. The third column shows the feature sets for which a lower performance
was obtained than for the given feature set. For each comparison a p-value is given in the fourth column. Significant differences are listed bold.

In table 2.3 the performances of three context feature sets are compared to the performance of the standard feature set. All context feature sets performed significantly better than the standard feature set ($p<0.0001$), yielding an increase in mean sensitivity between 3% and 5.9%. Best results were obtained for the variant in which all views were used ($F_{AV}$) for computing the context features. These results were significantly better than when using only the ipsilateral view. In Figure 2.5 the FROC curves are shown obtained for the feature sets $F_0$, $F_{IL}$ and $F_{AV}$. Vertical lines show the interval in which the mean sensitivity $S$ is computed. As described in the methods section, sensitivity was defined as the fraction of abnormal cases detected. When the sensitivity was computed as the fraction of abnormal mass regions detected, we found an increase in mean sensitivity between 2.6% and 6.2%. Similar to the case based performances, best results were obtained for the variant in which all views were used for computing the context features.

**Table 2.3:** Mean sensitivity $S$ for the standard feature set and for feature sets consisting of the standard features and the three context features computed using each view variant

<table>
<thead>
<tr>
<th>Feature set</th>
<th>$S$</th>
<th>Compared to</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$F_0$</td>
<td>0.604</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$F_{IL}$</td>
<td>0.634</td>
<td>$F_0$</td>
<td>$&lt;0.0001$</td>
</tr>
<tr>
<td>$F_{CL}$</td>
<td>0.655</td>
<td>$F_0$</td>
<td>$&lt;0.0001$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$F_{IL}$</td>
<td>0.0038</td>
</tr>
<tr>
<td>$F_{AV}$</td>
<td>0.663</td>
<td>$F_0$</td>
<td>$&lt;0.0001$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$F_{IL}$</td>
<td>$&lt;0.0001$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$F_{CL}$</td>
<td>0.0248</td>
</tr>
</tbody>
</table>

Table 2.4 shows results for feature sets in which one or two context features (AV-loc, AV-ima and/or AV-b) were added to the standard feature set. When comparing the feature sets in which only one context feature was added, $F_{ima}$ yielded the best performance. However, this performance was not significantly higher than the performance obtained when using $F_{loc}$ or $F_0$. On the other hand, a combination of any two of these features yielded a significant performance increase compared to $F_{loc}$. Best results were obtained when the features AV-ima and AV-b were combined. This performance
2.3 Results

was significantly higher than the use of only one context feature \( (F_{loc}, F_{ima} \text{ or } F_{b}) \), but not significantly higher than the performance obtained when using another combination of two features \( (F_{loc,ima} \text{ or } F_{loc,b}) \) or when all three features were present \( (F_{AV}) \). In Figure 2.6 the FROC curves are shown obtained for the feature sets \( F_{0}, F_{ima} \text{ and } F_{ima,b} \). For the standard feature set \( (F_{0}) \) and the best performing feature set \( (F_{ima,b}) \) we also computed separate FROC curves for the mammograms taken at the screening round prior to detection and for the diagnostic mammograms. These curves are shown in Figure 2.7 and Figure 2.8. For both prior and diagnostic mammograms there is a large increase in sensitivity (case and lesion based) when context features are used.

![Figure 2.5: Case-based FROC curves for the standard feature set \( F_{0} \) and the sets in which the three context features are added computed using the ipsilateral view \( (F_{IL}) \) and using all views \( (F_{AV}) \). Vertical lines show the interval in which the mean sensitivity \( S \) is computed.](image)

It is expected that the number of false positive candidate regions found in dense breasts is relatively larger than the number of false positive regions found in fatty breasts. To investigate variability in performance improvement among density groups, we computed a density class for each case using the method described by Karssemeijer. This yielded 4 density groups. For the mammograms taken during a screening round prior to detection the number of cases in class 1 (fatty) to class 4 (very dense) were resp. 76, 98, 32 and 2. For the diagnostic mammograms the number of cases in class 1 to 4 were resp. 136, 185, 93 and 14. For the normal mammograms the num-
Table 2.4: Mean sensitivity $S$ for feature sets in which 1 or 2 context features, computed using all views, are added to the standard feature set

<table>
<thead>
<tr>
<th>Feature set</th>
<th>$S$</th>
<th>Compared to</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$F_{loc}$</td>
<td>0.635</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$F_{ima}$</td>
<td>0.652</td>
<td>$F_{loc}$</td>
<td>0.0138</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$F_{b}$</td>
<td>0.1080</td>
</tr>
<tr>
<td>$F_{b}$</td>
<td>0.644</td>
<td>$F_{loc}$</td>
<td>0.0798</td>
</tr>
<tr>
<td>$F_{loc,ima}$</td>
<td>0.661</td>
<td>$F_{loc}$</td>
<td>$&lt;0.0001$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$F_{ima}$</td>
<td>0.0198</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$F_{b}$</td>
<td>0.0038</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$F_{loc,b}$</td>
<td>0.0894</td>
</tr>
<tr>
<td>$F_{loc,b}$</td>
<td>0.654</td>
<td>$F_{loc}$</td>
<td>$0.0002$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$F_{ima}$</td>
<td>0.3340</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$F_{b}$</td>
<td>0.0090</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$F_{loc,b}$</td>
<td>0.0010</td>
</tr>
<tr>
<td>$F_{ima,b}$</td>
<td>0.665</td>
<td>$F_{loc}$</td>
<td>$&lt;0.0001$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$F_{ima}$</td>
<td>$0.0010$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$F_{b}$</td>
<td>$&lt;0.0001$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$F_{loc,ima}$</td>
<td>0.1560</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$F_{loc,b}$</td>
<td>0.0208</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$F_{AV}$</td>
<td>0.2704</td>
</tr>
</tbody>
</table>
Figure 2.6: Case-based FROC curves for the standard feature set $F_0$, the set in which the feature AV-ima is added ($F_{ima}$) and the set in which both features AV-ima and AV-b are added ($F_{ima,b}$). Vertical lines show the interval in which the mean sensitivity $S$ is computed.
Figure 2.7: Case-based FROC curves for the mammograms taken during a screening round prior to detection of the cancer (P) and for the diagnostic mammograms (D). Curves are shown for the standard feature set $F_0$ and the set in which both features AV-ima and AV-b are added ($F_{ima,b}$).
Figure 2.8: Lesion-based FROC curves for the mammograms taken during a screening round prior to detection of the cancer (P) and for the diagnostic mammograms (D). Curves are shown for the standard feature set $F_0$ and the set in which both features AV-ima and AV-b are added ($F_{ima,b}$).
bers were resp. 900, 1239, 907 and 216. Because in the Netherlands only women in the age interval of 50-75 years are invited for screening, the number of cases present in the fourth density group was relatively small. Therefore, we combined the third and fourth group. In table 2.5 the mean sensitivity measure $S$ is shown for each of the density groups when the standard feature set $F_0$ is used and when the context feature set $F_{ima,b}$ is used. These results indicate that larger performance improvements are expected for dense breasts than for fatty breasts.

Table 2.5: Mean sensitivity $S$ for feature sets $F_0$ and $F_{ima,b}$ computed for several density groups

<table>
<thead>
<tr>
<th>Density</th>
<th>Prior</th>
<th>Diagnostic</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$F_0$</td>
<td>$F_{ima,b}$</td>
</tr>
<tr>
<td>1</td>
<td>0.331</td>
<td>0.332</td>
</tr>
<tr>
<td>2</td>
<td>0.293</td>
<td>0.385</td>
</tr>
<tr>
<td>3 + 4</td>
<td>0.331</td>
<td>0.471</td>
</tr>
</tbody>
</table>

2.4 Discussion and conclusion

We developed context features that represent suspiciousness of normal tissue and suggested that these features are beneficial in detection of malignant masses in mammograms. This suggestion is supported by our results. Feature sets containing context features performed significantly better than an existing feature set without context, yielding an increase in mean sensitivity of more than 6%. Further, we found a significant performance increase when multiple views were used for computing context features compared to when a single view was used. This may be due to three reasons. The first reason is that the estimation of normal tissue context might be more accurate when multiple views are combined. Secondly, the use of normal tissue in the contralateral breast yields a measure of asymmetry, which is also known to be used by radiologists in detecting breast cancer. Finally, sometimes a large area of the breast is affected by breast cancer. If this is the case, our automated method of defining normal reference tissue in the breast may in fact include multifocal cancer areas. Therefore, in such cases it might be more beneficial to use reference areas in the projections of the contralateral breast.

Our results suggest that for best results several reference areas should be taken into
account for each view. We evaluated context features that were computed using three areas: the local surroundings, the whole breast and a band of locations with similar distances to the nipple. Best performance was obtained when a combination was used of context features based on the whole breast (and pectoral muscle) and based on the distance matched band. Although no significant differences were found when comparing this performance to other combinations of two context features or to the use of all three context features, this combination performed significantly better than the use of any single context feature.

An advantage of the described context features is that no additional segmentation is required other than segmentation of the candidate mass. This is in contrast to several other studies, in which a reference area is segmented and morphological features are computed, such as contrast and circularity of the reference region. Morphological features can not always be computed because there is not always a relevant region present that can be adequately segmented. The context features described in this article can be computed for each desired reference area because they are based on suspiciousness scores that are available for each location in the mammogram.

A second advantage is that local features computed for a reference area, such as spiculation and gradient values, are intelligently combined into a suspiciousness score before they are used as reference in the final interpretation stage. Although the final classifier also combines features in an intelligent way, the training of this second classifier is often based on a limited set of suspicious candidate regions. In contrast to this, the initial stage classifier is trained under supervision using a large number of feature vectors sampled in normal images. Therefore, this initial classifier is expected to obtain more information about normal tissue than the final stage classifier. A third advantage is that the suspiciousness scores for all locations in the mammogram are already computed in the initial detection stage. Therefore, the computation of the context features is done using simple statistics and is very fast.

Bibliography


Use of normal tissue context in computer-aided detection of masses in mammograms


The effect of feature selection methods on computer-aided detection of masses in mammograms

R. Hupse and N. Karssemeijer

Original title: The effect of feature selection methods on computer-aided detection of masses in mammograms

Published in: Physics in Medicine and Biology 2010;55:2893-2904
Abstract

In computer aided diagnosis (CAD) research, feature selection methods are often used to improve generalization performance of classifiers and shorten computation times. In an application that detects malignant masses in mammograms, we investigated the effect of using a selection criterion that is similar to the final performance measure we are optimizing, namely the mean sensitivity of the system in a predefined range of the Free-response Receiver Operating Characteristic (FROC). To obtain the generalization performance of the selected feature subsets, a cross validation procedure was performed on a data set containing 351 abnormal and 7879 normal regions, each region providing a set of 71 mass features. The same number of noise features, not containing any information, were added to investigate the ability of the feature selection algorithms to distinguish between useful and non-useful features. It was found that significantly higher performances were obtained using feature sets selected by the general test statistic Wilks’ lambda than using feature sets selected by the more specific FROC measure. Features selection led to better performance when compared to a system in which all features were used.
3.1 Introduction

Computer aided detection (CAD) systems are being developed for a wide range of applications in radiology. Most notable are applications in mammography, lung CT, and virtual colonoscopy. These systems perform an independent automated interpretation of image data and detect potential abnormalities. By placing attention markers on detected regions when radiologists are interpreting the images it may be avoided that significant abnormalities are overlooked. In practice CAD systems operating according to this principle are already widely used in breast cancer screening and most clinical studies confirm that they are effective, despite the fact that their standalone performance is generally worse than that of the radiologists. In a recent prospective randomized trial it was found that single reading with CAD yielded similar results as double reading in a breast cancer screening program \(^1\).

The approach developers take to build computer aided detection systems is similar in many applications. In an initial detection stage candidate regions are identified based on local image features that are relatively inexpensive to compute. Subsequently, rich descriptions of the candidate regions are generated using sophisticated feature extraction techniques and supervised machine learning methods are applied to estimate the posterior probability that a region is a true abnormality given its features. Finally, a threshold on the posterior probability determines which regions are marked. To avoid that the system misses abnormalities a low threshold is used, which inevitably leads to many false positives. In practice, display of a few marks per case is accepted by most the readers as long as the sensitivity is high enough.

To develop CAD systems, availability of large databases with representative cases and reliable annotations of abnormalities is crucial. Increasing the number of cases for training of a CAD system makes it possible to extend the description of candidate regions detected in the initial stage, which leads to a gradual increase in performance. To determine whether extended region descriptions are useful, feature selection methods are used. By selecting a subset of features from a large pool of descriptors the performance of a system may be increased, because the risk of overfitting a classifier to the training data is reduced when less features are used. This is true particularly if the number of features is large in comparison to the size of the training data, which is often the case during development of a system, where one is exploring novel feature extraction approaches. Other benefits of feature selection are reduced computational demands, shorter training and validation cycles, and the possibility it offers to gain more understanding of the system and its remaining weaknesses. With huge databases becoming available for certain applications, there is a growing need to investigate effectiveness of feature selection methods. Additional information provided by novel
sets of features in an application that already performs well may be incremental but still significant and important.

Feature selection strongly depends on the choice of a performance measure used as optimization criterion. One would expect to obtain best results when an optimization criterion is used that reflects use of the CAD system in clinical practice. However, it is common CAD research to use more general feature selection criteria that do not directly reflect performance at operating points used in clinical practice. To evaluate CAD systems the methodology of Free Response Operating Characteristics (FROC) is widely used. Therefore, in this study we investigated use of an FROC based optimization criterion, which is directly related to the performance measure used in clinical practice. In particular, we investigated if performance in a predefined range of the FROC curve can be improved in this way. We compared this method to the standard approach of using Wilks’ lambda as optimization criterion in a CAD application aimed at detection of masses in screening mammograms. Using a very large database for training and validation, we also investigated the effect of feature selection on the quality of the final CAD performance in comparison to using all available features.

3.2 Methods

3.2.1 Data set and preprocessing

Mammograms were taken from an annotated database containing cases from the Dutch breast cancer screening program. From this database we randomly selected 213 cases containing a malignant mass. In this study we use the term ‘mass’ for the group of malignant masses, architectural distortions and focal asymmetries. The mass cases were a mixture of screen detected cancers (143 cases, 545 images) and priors (70 cases, 181 images) and contained at least one visible and biopsy-proven malignant mass. All mass regions were annotated under supervision of an experienced radiologist. A total of 336 normal cases (1122 images) were also randomly selected. We used approximately a third of all available data in the database. The reason for not using all data is that the effect of feature selection is larger when the dataset is smaller. In the Netherlands, two mammographical views (medial lateral oblique and cranio caudal) are obtained at the initial screening, and only one view (medial lateral oblique) on the subsequent screenings unless there is an indication that obtaining the second view would be beneficial. The percentage of abnormal cases for which both views were available was 70%, for normal cases this percentage was 67%. Part of the mammograms (274 cases) were digitized using a Lumisys 85 digitizer and the other part (275 cases) using a Canon digitizer. All mammograms were digitized at a pixel resolution of 50µm and were
sampled down to a resolution of 200μm. The gray value depth was 12 bits. Preprocessing consisted of three steps: segmentation into three areas (breast region, background and pectoral muscle), pectoral equalization and peripheral enhancement. These steps are described by Karssemeijer.  

### 3.2.2 Candidate region detection and feature calculation

For all images an initial candidate generation was performed as described by Karssemeijer and te Brake. Using this method five texture features were computed on a regular grid in the visible part of the breast and pectoral muscle. These texture features were computed using the intensity values of the surrounding area (see Karssemeijer and te Brake for more details). Three of these features were related to the presence of spicules, two were related to the presence of a central mass. For each location a suspiciousness level was calculated using five neural networks trained with different random initializations on a separate dataset. The averaged output of these neural networks was considered the suspiciousness level for the location, and all suspiciousness levels were combined into a suspiciousness image. Each spatial local maximum in the suspiciousness image that exceeded a (relatively low) threshold was considered a mass candidate, and for each candidate mass location a region was segmented using a segmentation method based on dynamic programming. In this way 351 true positive and 7879 false positive regions were segmented. A mass region was detected if the center of mass of the segmented CAD region was located inside the region that was annotated by the radiologist. Some mass regions were detected by more than one segmented CAD region. For each segmented region, a set of 71 features was calculated. The features can be subdivided in several categories. These categories are listed in table 3.1. One of the features groups measure the suspiciousness of normal tissue. These context features are described in. Besides the described features we assigned 71 random values to each candidate region. We refer to these random values as noise features. The distribution of these noise feature values was uniform between 0 and 1. Although these features were theoretically not useful for CAD, they were added to investigate the ability of the feature selection algorithms to distinguish between useful and non-useful features.

### 3.2.3 Feature selection methods

**Search algorithm**

We used the sequential floating forward selection (SFFS) algorithm as proposed by Pudil and modified by Spence and Sajda. In this algorithm addition and removal
The effect of feature selection methods on computer-aided detection of masses

Table 3.1: Features computed for each segmented candidate region

<table>
<thead>
<tr>
<th>Category</th>
<th>Number</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial detection</td>
<td>11</td>
<td>Local spiculation and mass measures, suspiciousness level, spiculation and mass measures for region</td>
</tr>
<tr>
<td>Normal tissue</td>
<td>17</td>
<td>Suspiciousness measures for normal tissue context</td>
</tr>
<tr>
<td>Location</td>
<td>10</td>
<td>Relative location and distances to skin, pectoral muscle, chest nipple</td>
</tr>
<tr>
<td>Shape and size</td>
<td>7</td>
<td>Acutance, compactness and measures of region size</td>
</tr>
<tr>
<td>Linear texture</td>
<td>5</td>
<td>Presence of linear texture in the region and its surround</td>
</tr>
<tr>
<td>Dense tissue</td>
<td>7</td>
<td>Features that determine the amount of dense tissue located inside and outside the region</td>
</tr>
<tr>
<td>Contrast</td>
<td>8</td>
<td>Difference between gray level distribution in the region and its surround</td>
</tr>
<tr>
<td>Border</td>
<td>6</td>
<td>Features measuring the continuity of the segmentation border</td>
</tr>
<tr>
<td>Total</td>
<td>71</td>
<td></td>
</tr>
</tbody>
</table>

of features is repeated alternately in a stepwise manner. In general, after addition of each feature there is a back-track loop in which features are removed from the set. The feature to be added or removed is always the feature that yields a better performance than all subsets obtained when one of the other features was added or removed instead. After adding or removing a feature, the chosen subset and its performance are stored in memory. Because features are added and removed alternately, the size of the subset increases and decreases during feature selection. A subset and its performance are only stored in memory if no subset of the given size is stored in memory yet, or if the new performance exceeds the performance of the subset of the same size currently stored in memory. In this way, performances are compared between subsets of
the same size. If adding a feature yields a performance worse than the performance of the stored subset, the algorithm continues using the stored subset. If removing a feature does not improve the performance of the subset in memory, the feature is not removed, the back-track loop stops and the algorithm continues with adding a feature. The algorithm stops if a predefined subset size is reached.

The selection algorithm described is very similar to the method called stepwise linear discriminant analysis (SWLDA) which is described in\textsuperscript{9–12}. In SWLDA features are also stepwise added and removed from the subset. The main difference is that in SWLDA the decision to add or remove a feature is based on the comparison between the performance of the obtained feature set and the performance of the previous feature set (in which the new feature is not yet added, or in which the feature is not yet removed). In this way, subsets of different sizes are compared. During feature selection, the data used for computing the performance measure is the same data as used for training the classifier. Therefore it is expected that a subset with a larger size will automatically result in a better performance. To make a decision, thresholds are used that indicate the minimal performance increase necessary to add or remove a feature from the set. The advantage of comparing subsets of the same size is that no thresholds are needed.

In contrast to SWLDA, the algorithm we use does not automatically stop at the best performing subset. Instead, it continues selecting features until a predefined subset size is reached. To obtain the single subset that performs best, one of the subsets can be chosen afterwards by comparing the performances during selection.

**Classifier**

Cross validation was used to validate the selected feature sets. The classifier we used during cross validation was a neural network ensemble that was trained separately from the feature selection process. However, the use of an ensemble of neural networks during feature selection was computationally too expensive because the classifier must be trained for each candidate feature set. Therefore we used a linear discriminant analysis classifier (Fisher’s LDA) during feature selection. A linear classifier is very practical because it can be computed fast. The Fisher’s discriminant score $d_i$ for a feature vector $y_i$ is computed by

$$d_i = w^T y_i + w_0$$  \hspace{1cm} (3.1)
with \( \mathbf{w} \) a vector of weights and \( w_0 \) a constant. The vector \( \mathbf{w} \) is found by maximizing the separability between both classes for data samples in feature space projected onto a line that points in the direction of \( \mathbf{w} \). The separability is measured by the Fisher criterion:

\[
J(\mathbf{w}) = \frac{|m_1 - m_2|^2}{\frac{1}{N_1} \sum_{i \in \text{class}1} (d_i - m_1)^2 + \frac{1}{N_2} \sum_{i \in \text{class}2} (d_i - m_2)^2}
\] (3.2)

with \( m_1 \) and \( m_2 \) the means of the discriminant scores and \( N_1 \) and \( N_2 \) the number of data samples for classes 1 and 2 respectively. If the feature values for both classes are normal distributed with equal covariance matrices, the application of Bayes rule results in a linear discriminant classifier that is equal to Fisher’s discriminant up to a constant\(^{13}\). This means that under these conditions Fisher’s discriminant is the optimal classifier. Because the actual classifier we use in our CAD system is a neural network ensemble that we train separately from feature selection, it is not necessary that the classifier used for feature selection is the optimal classifier. However, the classifier must perform well enough to select the features that perform best in the final classifier.

**Performance measures**

In each step of the selection procedure the feature has to be selected that yields the best performing subset compared to other candidate subsets. To select this feature, a robust performance measure is required. When using the Fisher’s discriminant classifier it would be straightforward and convenient to use the Fisher criterion to measure performance. However, in statistical pattern recognition it is more common to use a likelihood ratio statistic such as Wilks’ lambda\(^{14}\). Wilks’ lambda is also well known from its use in multivariate analysis. The statistic has a know distribution which allows computation of p-values, which are used in the SWLDA feature selection procedure. Wilks’ lambda is defined as the ratio of within-group sum of squares to the total sum of squares of the linear discriminant scores:

\[
\lambda = \frac{\sum_{i \in \text{class}1} (d_i - m_1)^2 + \sum_{i \in \text{class}2} (d_i - m_2)^2}{\sum_{i=1}^{N_{\text{total}}} (d_i - m)^2},
\] (3.3)

with \( d_i \) the discriminant score for region \( i \), \( m_1 \) and \( m_2 \) the means of the discriminant scores for classes 1 and 2 respectively, \( m \) the mean over both classes and \( N_{\text{total}} \) the total
number of regions. It can be seen that this measure is related to the Fisher criterion, because both use the within-group sum of squares. A small value for Wilks’ lambda means a large separation between classes.

The performance criterion outlined above measures global separation between the classes. However, in CAD applications one is often more interested in performance within a given range of operating points that corresponds to use of the system in practice. We determine the final test set performance of the CAD system using a measure based on FROC analysis. In our experiments, this FROC based measure was used as criterion during feature selection besides the use of Wilks’ lambda. We define the measure $SFROC$ as the mean sensitivity in a range of false positive levels computed on a logarithmic scale:

$$SFROC = \frac{1}{\ln(1.0) - \ln(0.1)} \int_{0.1}^{1.0} \frac{s(f)}{f} df,$$

with $f$ the number of false positives in normal images and $s(f)$ the lesion sensitivity. This measure is proportional to the partial area under the FROC curve plotted on a logarithmic scale (see figure 3.1). We chose the false positive range from 0.1 to 1.0 false positives/image as this corresponds best to the setting of current CAD systems in screening practice. Use of a logarithmic scale avoids that the measure is dominated by operating points at high false positive rates.

The FROC measure defined above has some important differences in comparison to Wilks’ lambda. First of all, the fact that the data we work with is heavily imbalanced with respect to the number of samples in both classes (there are much more false positives) has a strong impact on the value of Wilks’ lambda while it does not affect the FROC measure. Secondly, the FROC measure focuses on relatively strong false positives and is insensitive to changes in the criterion used to select the number of candidate regions after the initial stage of the CAD system. Therefore, we investigated if using the FROC criterion $SFROC$ for feature selection has advantages over the use of Wilks’ lambda.

### 3.2.4 Cross validation

Feature selection was performed as part of a 5-fold cross validation. In each cross validation round 4/5th of the cases were used for selecting features. Feature selection was automatically stopped when a subset size of 125 was reached. After feature selection a final performance measurement was obtained for selected subsets of size 1, 2,
The effect of feature selection methods on computer-aided detection of masses

Figure 3.1: An FROC curve. The measure $SFROC$ is defined as the mean sensitivity in a range of false positive levels computed on a logarithmic scale. We chose to use a range from 0.1 to 1.0 false positives/image.

3, 4 and 5, 10-40 in steps of 5 and 60-120 in steps of 20. This was done by applying a classifier on the other 1/5th of the dataset. The classifier was either an LDA classifier (the same classifier as used during feature selection), or a neural network classifier. The neural network classifier consisted of an ensemble of five neural networks. Each network was initiated with different random weights and trained using the selected feature set. Training was done using the back-propagation algorithm. Each network consisted of an input layer, a hidden layer of 12 nodes and an output layer of one node. A separate data part (that was not part of the cross validation data set) was used for determining a learning curve during training. The training of a network stopped when the performance of the network on this separate data set was maximal. The output of the neural network ensemble was defined as the average output of the five networks. Because the ratio true positives versus false positives in the training sets was very low (approximately 1:22), we decided to use a fixed ratio of 1:9 in which the true positives and false positives were presented to the network. The choice of this ratio was based on former experiences.

After cross validation we pooled the test results for all cases and two FROC curves were constructed, one for the lesion-based sensitivity and one for the case-based sensitivity. Lesion-based sensitivity was defined as the fraction of annotations being detected, case-based sensitivity was defined as the fraction malignant cases being detected. A case was considered detected when at least one annotated malignant region
present in the case was detected. When constructing the FROC curve, false positives were only counted on the images of normal cases. The final performance measure for each feature subset was $SFROC$, computed by eq. 3.4. For comparison, the cross validation procedure was also performed without feature selection, using the entire set of available features for training and testing.

### 3.2.5 Performance and statistical analysis

We investigated differences in performance between same sized subsets selected using Wilks’ lambda as performance measure and subsets selected using $SFROC$ as performance measure. Performance differences were also computed between selected subsets and the feature set containing all features. A confidence interval of the performance difference was computed using the bootstrap method. By bootstrapping, the performance measure $SFROC$ was obtained a large number of times for two feature sets, however, each time computed using a resampled data set. Every bootstrap sample contained the same number of cases as the total image set. These cases were sampled from the total set of cases with replacement. For each sampled data set two FROC curves were computed, one for each feature set. For these two FROC curves the difference in $SFROC$ was calculated. The sampling was repeating 1000 times and resulted in 1000 difference values for $SFROC$. A 95% confidence interval for this difference was obtained by sorting the difference values and setting the 25th and 975th value as under and upper limit.

### 3.3 Results

Figure 3.2 shows the cross validation performance obtained when using the neural network classifier for validation of the feature sets selected using the $SFROC$ criterion and the Wilks’ lambda criterion. For both feature selection methods, the performance measure $SFROC$ was computed for each selected subset using the lesion-based FROC curve (a) and the case-based FROC curve (b). The performance obtained when all features were used to train the neural network classifier is also shown. In (c) the difference is plotted between the lesion-based performance obtained using the subsets selected by Wilks lambda and the set containing all features. For this difference the 95% confidence interval is given. This interval is completely above zero for subset sizes 15-30 and 40. Figure 3.2(d) shows the lesion-based performance difference obtained for the subsets selected using Wilks’ lambda and the subsets selected using the FROC measure. The confidence interval of this difference is completely above zero for subset sizes 1 and 10-100. Confidence intervals for case-based performance differences were also computed.
The effect of feature selection methods on computer-aided detection of masses

(not plotted). When comparing Wilks’ lambda and the set containing all features, the confidence interval of the case-based performance difference is complete above zero for subset sizes 30 and 40. When comparing Wilks’ lambda and the FROC measure, the confidence interval of the case-based performance difference is completely above zero for subset sizes 1, 15 and 30-100.

![Figure 3.2](image)

**Figure 3.2:** Lesion-based (a) and case-based (b) performances obtained for feature sets selected using the FROC and the Wilks’ lambda criterion. Performances were obtained using a neural network classifier. The performance obtained when all features were used to train the neural network classifier is also shown. The difference in lesion performance between the subsets selected by Wilks’ lambda and the set containing all features is shown in (c). The difference in lesion performance between the subsets selected by Wilks’ lambda and the subsets selected by the FROC measure is shown in (d). For each performance difference, the 95\% confidence interval is given.

In figure 3.3 cross validation results are shown which were obtained using an LDA classifier instead of the neural network classifier. For comparison, the result of training the neural network classifier with all available features is also plotted. Figure 3.3(a) shows the lesion-based performance measures, figure 3.3(b) shows the case-based per-
3.3 Results

Performance measures. Differences and confidence intervals between the obtained performances were computed (not plotted). When comparing the lesion-based performances obtained using the SFROC and Wilks’ lambda selection criterion, for none of the subset sizes a significant difference in performance was found. When comparing the case-based performances, the 95% confidence interval of the performance difference obtained using the SFROC and Wilks’ lambda selection criterion is completely above zero for subset size 80. For the lesion-based performances, none of the selected subsets yielded significantly better results than the set with all available features. For the case-based performances, only the subsets of size 80 and 100, selected with the SFROC criterion yielded a significant improvement compared to the set with all features.

![image](a)

**Figure 3.3:** Lesion-based (a) and case-based (b) performances obtained for feature sets selected using the FROC and the Wilks’ lambda criterion and validated using an LDA classifier. For comparison, the result of training the neural network classifier with all available features is also plotted.

We also computed the 95% confidence interval between the performance difference obtained for the complete feature set in the case of using a neural network classifier and in the case of using an LDA classifier. This interval was [-0.002, 0.051] for the lesion-based performance, and [0.006, 0.070] for the case-based performance.

Figure 3.4 shows the average overlap of features that were selected using a feature selection method. The overlap was defined as the average number of identical features present in two subsets of the same size, selected by the same feature selection method. We computed this by averaging the number over each combination of 2 sets selected during the 5 cross validation rounds. Overlap is shown for each subset size. Noise features were not counted when computing the overlap. The numbers of selected non-noise features (averaged over the 5 cross validation rounds) are also plotted.
The effect of feature selection methods on computer-aided detection of masses

Figure 3.4: The overlap between feature sets chosen during the cross validation rounds and the average number of non-noise features, plotted for each subset size and feature selection method.

Results show that there is more overlap between the subsets selected by Wilks’ lambda than by the subsets selected by the FROC measure. Further, less noise features were chosen when using Wilks’ lambda.

3.4 Discussion and conclusion

We compared different feature selection methods to select subsets from a set of 142 features. One half of this feature set was extracted from candidate mass regions detected in a large set of mammograms. The other half were noise features, not containing any information about the regions. After feature selection, we used either a neural network classifier or an LDA classifier for validation of the chosen subsets.

Best performances were obtained using the neural network classifier for validation. In this case, for most sizes of selected subsets the subsets selected using the general test statistic Wilks’ lambda as selection criterion performed better than subsets selected using a more specific measure defined as the mean sensitivity in a specificity interval of the FROC curve. This difference was significant for both lesion-based and case-based performance. For some subsets selected by Wilks’ lambda, the performance was also significantly better than the performance obtained using the complete feature set of 142 features. This performance increase was not obtained using the FROC measure as
The smallest subset we tested that yielded a significant increase in lesion-based performance compared to the complete feature set consisted of only 15 features. For the case-based performance a subset of 30 features yielded a significant performance increase. Although the performance increase was small (approximately 2 percent), the number of features needed to achieve this performance was much smaller than when using the complete feature set. One of the reasons that the increase in performance was not very large might be that we used a training schedule for the neural network classifier that minimizes the risk of overtraining. This was done by using a separate validation data set to compute a training learning curve and automatically stopped training when the performance on this data was maximal. Further, by using a relatively small number of hidden nodes in the network overtraining might also be avoided.

During feature selection, always an LDA classifier was used. The use of a neural network classifier during feature selection was computationally too expensive because the classifier must be trained for each candidate feature set. For validation of the selected feature sets two types of classifiers were used, a neural network classifier and an LDA classifier. It was found that the performance of the complete feature set decreased for both lesion-based and case-based measures when an LDA classifier was used instead of a neural network classifier. Bootstrapping indicated that the difference in case-based performance was significant. Probably some relations between feature values and the class label can only be learned by a more complex classifier like the neural network. When the LDA classifier was used for validation of the selected feature sets, only two feature subsets yielded a performance that was significantly better than the set with all features. Both were chosen using the SFROC criterion. For most subsets chosen using the SFROC criterion, case-based performances were higher than for subsets chosen using the Wilks’ lambda criterion. However, this difference was significant for only one subset size. It seems that the SFROC criterion selects features which give relatively best results when used in an LDA classifier. However, when a neural network classifier is used, better performances are obtained with feature sets selected using the Wilks’ lambda criterion.

Overall, best results were obtained when Wilks’ lambda was used as selection criterion during feature selection and the final classifier was trained using a neural network classifier. This is in contrast to our expectation that the best results would be obtained when a performance measure is used during feature selection that is similar to the final validation performance measure to be optimized. Our results suggest that Wilks’ lambda is able to select features that are more general to the data than the FROC based performance measure. A reason might be that relatively few false positive mass regions get a suspiciousness level assigned that is high enough to influence the
The effect of feature selection methods on computer-aided detection of masses

sensitivity in the chosen specificity interval of the FROC curve. Significant differences in suspiciousness levels are more difficult to detect in such a small number of regions due to random effects. Although we are more interested in differences in sensitivity for high specificities, it seems that a general statistic as Wilks’ lambda is more powerful for selecting a good feature set than the criterion \( SFROC \) because the FROC curve in a high specificity interval is too noisy. This idea is supported by the lower number of noise features chosen by Wilks’ lambda and the larger overlap between the subsets for different cross validation rounds.

Bibliography


Using computer aided detection in mammography as a decision support

M. Samulski, R. Hupse, C. Boetes, R.D. Mus, G.J. den Heeten and N. 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.
4.1 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 \(^1^–^3\). Nevertheless, there is a continuing debate regarding the usefulness of CAD\(^4^,^5\). 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\(^6^–^8\). 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\(^9\). 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\(^10^–^12\). Therefore, it is not surprising that studies reveal that many significant lesions are still missed even when CAD marks them\(^13^–^16\). 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\(^17^,^18\), 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\(^10\). In that study, CAD marks on regions not reported by the reader were not used, which is similar to the approach investigated here. As independent combination of reader results with CAD would not be easily accepted
Using computer aided detection in mammography as a decision support

Figure 4.1: 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.

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.

4.2 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
4.2 Materials and methods

× 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 4.1) 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\textsuperscript{19–22}.

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.

4.2.1 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 \( \mu m \). The mammograms were averaged down to a resolution of 100 \( \mu m \), 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 4.1 the study is summarized.
Table 4.1: Study overview

<table>
<thead>
<tr>
<th>Total cases</th>
<th>120</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal cases</td>
<td>80</td>
</tr>
<tr>
<td>Cancer cases</td>
<td>40</td>
</tr>
<tr>
<td>Cancer cases detected by CAD( ^a )</td>
<td>33</td>
</tr>
<tr>
<td>Available CAD regions( ^b )</td>
<td>587</td>
</tr>
<tr>
<td>Available true-positive CAD regions</td>
<td>41</td>
</tr>
<tr>
<td>Available false-positive CAD regions</td>
<td>546</td>
</tr>
</tbody>
</table>

\( ^a \) 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

4.2.2 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.

4.2.3 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. 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.

4.2.4 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.

4.2.5 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.

4.3 Results

The results of the nine individual readers are shown in Table 4.2. 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 4.2, 4.3 and 4.4, 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

66
Table 4.2: Reader detection performance in the false-positive fraction interval ranging from 0 to 0.1

<table>
<thead>
<tr>
<th></th>
<th>Without CAD TPF10 (%)</th>
<th>With CAD TPF10 (%)</th>
<th>Independent combination TPF10 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-radiologists</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>41.1</td>
<td>51.3</td>
<td>43.3</td>
</tr>
<tr>
<td>2</td>
<td>35.3</td>
<td>51.5</td>
<td>41.7</td>
</tr>
<tr>
<td>3</td>
<td>16.0</td>
<td>25.9</td>
<td>26.3</td>
</tr>
<tr>
<td>4</td>
<td>15.4</td>
<td>25.2</td>
<td>27.4</td>
</tr>
<tr>
<td>5</td>
<td>18.3</td>
<td>41.9</td>
<td>26.7</td>
</tr>
<tr>
<td>Average</td>
<td>25.2</td>
<td>39.2</td>
<td>33.0</td>
</tr>
<tr>
<td>Radiologists</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>24.3</td>
<td>32.3</td>
<td>33.6</td>
</tr>
<tr>
<td>7</td>
<td>24.8</td>
<td>28.8</td>
<td>30.2</td>
</tr>
<tr>
<td>8</td>
<td>30.2</td>
<td>25.7</td>
<td>37.0</td>
</tr>
<tr>
<td>9</td>
<td>20.2</td>
<td>30.4</td>
<td>30.0</td>
</tr>
<tr>
<td>Average</td>
<td>24.9</td>
<td>29.3</td>
<td>32.7</td>
</tr>
<tr>
<td>Reader average</td>
<td>25.1</td>
<td>34.8</td>
<td>32.9</td>
</tr>
</tbody>
</table>
Figure 4.2: 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 4.3: Average LROC curves obtained from the five non-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 4.5. 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 4.6, the same case is shown with the activated CAD region. The average time to read a case without CAD was \(84.7 \text{ seconds} \pm 61.5\). The radiologists read the cases much faster than the non-radiologists. Average reading time in the session with CAD was \(85.9 \text{ seconds} \pm 57.8\) per case (Table 4.3). There were no significant differences in reading times for the session with CAD and the session without CAD (\(p = 0.13\)) (Table 4.3). 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 4.4 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.
Using computer aided detection in mammography as a decision support

Figure 4.5: 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.
4.3 Results

Figure 4.6: The same case as in Figure 4.5 with the activated CAD region. The red contour and a CAD score close to zero indicate a high probability that this is a cancer.
4.4 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. 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. 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
4.4 Discussion

Table 4.4: Number of CAD regions queried\textsuperscript{a}

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

\textsuperscript{a} There were 587 CAD regions in total; 546 false-positive CAD regions and 41 true-positive CAD regions
analysis was restricted to locations identified by the radiologists\textsuperscript{9}. 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\textsuperscript{1} 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\textsuperscript{10}. Results in this study confirm that by independent combination of reader scores with CAD performance can be improved (Table 4.2). 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 4.3 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 for financial and technical reasons a 4-megapixel color display was used, 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. 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. 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
Using computer aided detection in mammography as a decision support

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\textsuperscript{17}. 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 to 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.

4.5 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.

Bibliography


Using computer aided detection in mammography as a decision support


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


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

Published in: Radiology 2012 (in press)
Abstract

Purpose: To compare effectiveness of an interactive computer-aided detection (CAD) system, in which CAD marks and their associated suspiciousness scores remain hidden unless their location is queried by the reader, to the effect of traditional CAD prompts used in current clinical practice for the detection of malignant masses in full field digital mammograms.

Materials and methods: The requirement for IRB approval was waived for the conducted retrospective observer study. Nine certified screening radiologists and three residents trained in breast imaging twice read 200 screening cases (63 with screen-detected malignant masses, 17 false-negatives, 20 false-positives, and 100 normals), once with CAD prompts and once with interactive CAD. Localized findings were reported and scored by the readers. In the prompting mode, findings were recorded before and after activating CAD. The partial area under the LROC curve for an interval of low false-positive fractions typical for screening (0 - 0.2) was computed for each reader and each mode. Differences in reader performance were analyzed using the DBM-MRMC method.

Results: Averaged over all readers, the partial area under the LROC curve for unaided reading was 0.57 and increased to 0.62 with interactive CAD, while remaining unaffected by prompting. The difference in reader performance for unaided reading and interactive CAD was significant (p=0.009).

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.
5.1 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 \(^1\text{-}^4\) but other studies did not demonstrate performance increase with CAD \(^5\text{-}^8\). In general, most radiologists agree that CAD is helpful for detection of microcalcifications, for which the sensitivity of CAD is high. However, there is less agreement about the benefit of CAD for detection of masses and architectural distortions. Many radiologists argue that CAD has too many false-positives to have a positive effect on mass detection \(^9\text{-}^{10}\).

Disappointing results of CAD may be due to the fact that masses are often missed because of incorrect interpretation \(^11\text{-}^{12}\), which is not the focus of existing CAD technology. Interestingly, in previous research it was found that reader performance can be improved with CAD by simply combining reader scores with the presence and probability of CAD mass markers \(^13\text{-}^{14}\). These results motivated us to develop a CAD system aimed at aiding radiologists with interpretation of suspicious regions rather than helping them with their initial detection. In the interactive system we propose, CAD marks are only displayed on demand for queried regions, together with a suspiciousness score.

The proposed system differs from other interactive CAD systems that have been developed, which provide additional information to justify CAD marks, in order to avoid that radiologists ignore them, or are intended to aid with lesion characterization in a clinical setting using reference libraries of similar cases \(^15\text{-}^{17}\). Our system is aimed at screening, but instead of justifying marks provided in the existing CAD paradigm, it relies completely on the reader for the initial detection process and only aims at improving interpretation and recall decisions. In a recent study, this novel way of using CAD for detection of masses was found to be effective \(^18\), however, without comparing it to the regular use of CAD prompts. In that study digitized film mammograms were used. The purpose of this study is to compare effectiveness of an interactive computer-aided detection (CAD) system, in which CAD marks and their associated suspiciousness scores remain hidden unless their location is queried by the reader, to the effect of traditional CAD prompts used in current clinical practice for the detection of malignant masses in full field digital mammograms.
5.2 Materials and methods

5.2.1 Study population

This retrospective study has been carried out in accordance with the applicable rules in the Netherlands concerning the review of research ethics committees and informed consent. All material was anonymized and 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 in Utrecht, the Netherlands<sup>19</sup>. In the screening program, women in the age group 50-74 are invited to participate every two years. Digital mammograms were acquired with a Selenia system (Hologic, Danbury, CT). All mammograms were read independently by two radiologists, with recall based on consensus. For subsequent screenings, digitized prior film mammograms were available of the exam preceding the first digital screening exam. The material from which we selected cases for the study included over 1200 recalls in which 202 cancers were detected with a mass or architectural distortion as the dominant sign of abnormality. Abnormalities were annotated under supervision of a radiologist.

5.2.2 Case selection

The screening pilot study<sup>19</sup> from which we collected material was a prospective study on the effect of digital screening. The purpose of the current study is the investigation of the use of CAD which is different from the pilot study in which the effect of CAD is not addressed. For the current study, we limited the number of cases to 200 because of practical consideration. We selected 80 biopsy-proven cancer cases and 120 negative cases because previous observer studies showed this was a reasonable balance. The cases were selected as follows. First, cases in which the lesion was rated as obvious, cases with only microcalcifications, and cases in which not all four views, i.e. the CC and MLO views of both breasts, were available were excluded. Then we checked for digital screening mammograms acquired prior to detection in which a malignant lesion was already visible. This yielded a total of 17 mammograms. From the remaining cases we randomly selected 63 screen-detected cancer cases from incident screening rounds. For the negative cases, we included 20 false positives verified by normal follow-up to make the study series more challenging. In these twenty cases, a suspicious mass or architectural distortion had been reported in screening, while biopsy was not deemed necessary during diagnostic assessment and recall was not due to a known benign abnormality such as cyst. In this way we sampled suspicious findings due to projection shadows that appeared normal during assessment. The remaining 100 negative mam-
mograms were randomly selected from the non-referred digital mammograms with at least one normal follow-up screening 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.

5.2.3 CAD and reading environment

The CAD system we developed for use in this study\textsuperscript{20,21} 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 preprocessing module this system can be used for full field digital mammograms\textsuperscript{22}. A special feature of the CAD system is that it automatically links regions in the MLO and CC views of the same breast if they correspond to the same lesion\textsuperscript{20,23–25}. This linking information is used when computing the suspiciousness score for a region.

The reader study was performed using an in-house developed experimental reading environment for screening mammography\textsuperscript{18}, 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 (FlexScan SX3031W; Eizo, Ishikawa, Japan) with a native resolution of 2,560 by 1,600. Actions of the readers are logged by the system to facilitate detailed analysis of the sessions.

CAD results were viewed in two different modes: the traditional prompting mode and the interactive mode. In the prompting mode, once activated all CAD regions were shown by displaying their contours, without providing a suspiciousness score. Prompts were shown for CAD regions with a suspiciousness score above a threshold. This threshold was adjusted in such a way that on reference set of normal mammograms on average 2 prompts were given per case. This prompting mode with the threshold we used is similar to use of CAD in current clinical practice.

In the interactive mode, regions detected by CAD remain hidden until activated by the reader, who can probe for CAD results with a mouse click 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, also for this other region the contour and score are shown.

To display a CAD result in the interactive mode, 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 four-view mammogram. More CAD results are accessible in this way than in the prompting mode, which is an inher-
ent advantage of the interactive system. We did not provide all marks accessible with interactive CAD also as prompts in the traditional reading mode, because this would not be tolerated by the readers who already complain about too many false positives, which might undermine our intention to compare our novel approach to CAD use in current clinical practice. In the interactive mode, contours of queried regions are 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 ranging from 0 (not suspicious) to 100 (very suspicious).

5.2.4 Observer study design

Twelve readers (nine radiologists with 1-24 years of experience in mammography and three residents trained in breast imaging) participated in the study. Before the study sessions, readers were offered a short training session to become familiarized with the experimental setup. Most readers had experience in using conventional CAD in screening practice. Five of the radiologists participated in earlier studies on our CAD system using traditional prompts. Three radiologists (reader 5, 8 and 11) had experience with the interactive display by participating in the prior study on interactive CAD. One reader was involved in reading part of the original screening mammograms which was more than 2 years before we conducted the observer study. Two readers were involved in the supervision of annotating the abnormalities which was done more than 1 year before conducting the observer study.

In the actual observer study, each reader read all cases in both modes in two sessions. In the first session, the first 100 cases were read with either prompting CAD or interactive CAD, while the second series of 100 cases were subsequently read in the alternate mode. In the second session, conducted at least 4 weeks after completing the first, the same cases were read again but with the reading modes swapped.

To obtain sufficient data for analysis, radiologists were asked to report more findings than they would normally do in screening practice. Readers marked the location of findings (not the contour) on both views and assigned a suspiciousness score in the range 0-100 to each finding. In the prompting mode, readers scored each case without CAD first, after which CAD was made available and readers could adjust scores and add new findings. Thus, results for three reading modes were obtained: unaided, regular CAD, and interactive CAD.
5.2.5 Data analysis

The standalone performance of CAD was computed using Free-response Receiver Operating Characteristic (FROC) analysis, providing the case-based true-positive fraction (TPF) as a function of the false-positive rate (FPR) in non-cancer cases. The location of a finding was considered correct if its distance to the center of the reference standard was less than 2 cm. In one case, two malignant masses were present. The correctly localized finding with the highest score was used for this case. For comparison we determined the sensitivity and specificity of the commercial R2 ImageChecker (V1.4, Hologic, Danbury, CT) on the study set. This system was adjusted to a specific setting.

Reader performance was computed using Location Receiver Operating Characteristic (LROC) analysis, which determines the TPF for each false-positive fraction (FPF), computed from the finding with the highest score in each of the non-cancer cases. For each reader and reading mode, the partial area under the LROC curve (pAUC) for $\text{FPF} < 0.2$ was computed. This was done using linear interpolation between the operating points. The low false-positive range was chosen to match operating points of radiologists in screening. In the dataset we used, 16.7% of the normal cases (20 of a total of 120) were recalled by the original screening radiologists. The raw LROC curves for each mode were averaged over all readers by computing TPF values for a standard set of FPF values using linear interpolation. This was done for FPF values that were reached by the LROC curves of all readers. Statistical analysis on the performance differences for the three modes was performed by the DBM-MRMC 2.3 method, which treats both readers and data as random samples using the jackknife method. This method does not take the location of findings into account. We consider treatment differences with a p-value smaller than 0.05 to be significant only if the global test on the null hypothesis of equal treatments is also significant (p-value < 0.05).

In the interactive mode the average number of queries with and without CAD response was computed in normal cases. Because in screening most cases are normal, this number reflects the number of clicks to be expected in screening practice. We also computed the median reading time for the normal cases for each reader. Median reading times were computed rather than average reading times, because the median is less affected by excessively long reading times caused by interruptions during the sessions. To investigate a potential effect of the experience of the readers on the benefit of CAD, the Pearson correlation coefficient was computed between the number of years readers practiced as qualified breast imager and the increase in pAUC with interactive CAD.
Figure 5.1: FROC curve for the CAD regions used in the study, computed over the 200 cases. The operating points for the traditional and the interactive mode are shown.

5.3 Results

The performance of CAD on the 200 study cases is shown in figure 8.1. In the prompting mode, the standalone sensitivity of CAD was 84% (67 of 80 cases detected) with on average 3.2 false-positives per normal case. In the interactive mode, the sensitivity was 91% (73 of 80 cases) with on average 8.2 false-positives per negative case. More false-positive marks were available than expected (2.0 for prompting and 8.0 for interactive CAD). The sensitivity of the commercial system was 0.75 at 0.21 FP/image. The sensitivity of our system used in the study was 0.76 at the same FPR.

In total 10031 findings were reported by the 12 readers in the 3 modes. By considering findings with a distance of more than 2 cm as unique, on average 958 unique findings were reported for each mode. All abnormalities were correctly localized by at least one reader in the unaided mode. Figure 8.2 shows the LROC curves for the three modes averaged over the readers. The partial area under the LROC curve (pAUC) is listed for each reader in table 5.1. For each reader the endpoint of the LROC curve was at a FPF value that was higher than the threshold used for computing the pAUC value (0.2), therefore no extrapolation was needed to compute the pAUC values. For 9 out of 12 readers the use of interactive CAD yielded a higher performance than unaided reading. On average the pAUC increased from 0.57 (SD: 0.088) to 0.62 (SD: 0.051). The difference in reader performance for these two modes was significant (p=0.009, with a p-value of 0.015 for the global test on the three modes). The average pAUC obtained with prompting was 0.57 (SD: 0.087), which was not better than unaided read-
5.4 Discussion

Reader performance with prompting was significantly worse than reading with interactive CAD (p=0.016). For three readers the performance was lower when interactive CAD was used compared to unaided reading. These were readers who achieved a relatively high performance in the unaided mode.

The correlation between the number of years with experience in mammography as a radiologist and the performance increase with interactive CAD is shown in figure 8.3. The correlation coefficient is -0.53, which suggests that performance increase is less for more experienced readers. The average number of clicks per normal case is listed in table 5.2 and the median of the reading times is listed in table 5.3. It appears that two of the most experienced readers (19 years of experience) probed regions extremely often, leading to display of most of the available CAD marks. Their performance dropped compared to unaided reading.

5.4 Discussion

We found that reader performance significantly increased when CAD results were interactively displayed. Prompting had no significant effect on reader performance. Most readers commented that they preferred the interactive system. The main reason

Figure 5.2: Location receiver operating characteristic (LROC) curves for unaided reading, with prompting CAD and with interactive CAD. The curves are averaged over all 12 readers.
Table 5.1: Partial area under the curve (pAUC) for false-positive fractions lower than 0.2 for the three modes and years of experience in mammography as a radiologist

<table>
<thead>
<tr>
<th>Reader</th>
<th>Unaided</th>
<th>Prompting vs. unaided</th>
<th>Interactive vs. unaided</th>
<th>Years experience</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.44</td>
<td>0.00</td>
<td>+0.11</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0.46</td>
<td>+0.01</td>
<td>+0.17</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0.52</td>
<td>+0.01</td>
<td>+0.04</td>
<td>14</td>
</tr>
<tr>
<td>4</td>
<td>0.53</td>
<td>0.00</td>
<td>+0.06</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>0.55</td>
<td>+0.02</td>
<td>+0.10</td>
<td>10</td>
</tr>
<tr>
<td>6</td>
<td>0.55</td>
<td>+0.01</td>
<td>+0.12</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>0.55</td>
<td>0.00</td>
<td>+0.01</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>0.56</td>
<td>−0.01</td>
<td>+0.04</td>
<td>20</td>
</tr>
<tr>
<td>9</td>
<td>0.61</td>
<td>0.00</td>
<td>−0.02</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>0.62</td>
<td>+0.01</td>
<td>+0.01</td>
<td>19</td>
</tr>
<tr>
<td>11</td>
<td>0.70</td>
<td>−0.02</td>
<td>−0.02</td>
<td>19</td>
</tr>
<tr>
<td>12</td>
<td>0.75</td>
<td>+0.01</td>
<td>−0.04</td>
<td>19</td>
</tr>
<tr>
<td>Average</td>
<td>0.57</td>
<td>0.00</td>
<td>+0.05</td>
<td>9</td>
</tr>
</tbody>
</table>

Note.- For the unaided mode the pAUC is given, for the other modes the increase in pAUC is given compared to the unaided mode. For the trained residents, the number of years with experience is set to zero.
5.4 Discussion

Figure 5.3: Correlation between the number of years with experience in mammography as a radiologist and the change in performance when interactive CAD is used. The Pearson correlation coefficient is -0.53.

Table 5.2: Number of clicks per normal case with and without CAD response

<table>
<thead>
<tr>
<th>Reader</th>
<th>With CAD response</th>
<th>Without CAD response</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.2</td>
<td>0.4</td>
<td>0.7</td>
</tr>
<tr>
<td>2</td>
<td>0.6</td>
<td>0.9</td>
<td>1.5</td>
</tr>
<tr>
<td>3</td>
<td>1.7</td>
<td>3.0</td>
<td>4.7</td>
</tr>
<tr>
<td>4</td>
<td>1.1</td>
<td>1.3</td>
<td>2.4</td>
</tr>
<tr>
<td>5</td>
<td>2.2</td>
<td>3.8</td>
<td>6.0</td>
</tr>
<tr>
<td>6</td>
<td>0.2</td>
<td>0.2</td>
<td>0.3</td>
</tr>
<tr>
<td>7</td>
<td>2.1</td>
<td>6.1</td>
<td>8.2</td>
</tr>
<tr>
<td>8</td>
<td>1.1</td>
<td>2.1</td>
<td>3.2</td>
</tr>
<tr>
<td>9</td>
<td>0.5</td>
<td>1.0</td>
<td>1.4</td>
</tr>
<tr>
<td>10</td>
<td>1.0</td>
<td>3.2</td>
<td>4.2</td>
</tr>
<tr>
<td>11</td>
<td>4.0</td>
<td>9.1</td>
<td>13.1</td>
</tr>
<tr>
<td>12</td>
<td>4.7</td>
<td>13.5</td>
<td>18.2</td>
</tr>
<tr>
<td>Average</td>
<td>1.5</td>
<td>3.8</td>
<td>5.3</td>
</tr>
</tbody>
</table>
may be that in the proposed interactive system the reading process is not disrupted by appearance of false positive prompts at unexpected locations, as is the case in the conventional system. 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. Suspiciousness scores of CAD generally correspond well with the observers own interpretation. When this is not the case, readers are alerted and pay more attention, which may lead to better decisions on average.

There is a large variance in the effect of interactive CAD for the 12 readers. Results suggest a trend that readers with more experience have a higher unaided performance and less or no benefit of CAD. For the two readers that probed regions extremely often, the decrease in performance may be related to this deviating use of CAD.

To make a fair and relevant comparison, the threshold for displaying prompts in the conventional system was set to a level corresponding to that in current clinical practice. In the study series, we measured false positive rate of 3.2 mass marks per case, which is in the range of settings used in commercial systems, but slightly higher than we expected. The reason for this was that the thresholds on the classifier output were based on a reference set of digitized film mammograms. It appeared that, on
5.4 Discussion

average, CAD scores for digital images were slightly higher than for digitized film. To be valid as a comparison to clinical practice, also the sensitivity of CAD should be high enough. Compared to the commercial R2 ImageChecker, the sensitivity of our system was slightly higher at the same FPR.

The threshold for determining which CAD results were available in the interactive system was different from the threshold chosen for displaying prompts in the conventional system, for reasons described in the methods section. We do not know if other results would have been obtained using another threshold. A further study is necessary to determine the optimal threshold.

In this study the effect of prompting may be underestimated. Our results differ from results obtained in prospective studies in which prompting had a positive effect on reader performance\(^1\)-\(^4\). It might be that in a retrospective observer study like ours less search errors are made and that therefore prompting has less effect. Another reason for the lack of an effect of regular CAD might be that a sequential reading order design was used to compare it to unaided performance. Such sequential reading approach has been criticized as it might bias reader performance for one of the two modes (unaided reading or reading with prompts). However, a study by Beiden\(^27\) shows little difference between mean effects of CAD measured in independent and sequential approaches. Further, reading with interactive CAD yielded an average reader performance that was significantly higher compared to both unaided reading and reading with prompts.

We found that the use of CAD (interactive or prompting) lengthened the reading time by approximately 10 seconds per case. It is noted that due to the sequential scoring of each case, the reading time with prompting could only increase compared to the unaided mode. Some readers reported that they spent more time exploring the CAD results in the interactive system out of curiosity, which will have increased the reading time. Therefore, we expect that with more experience reading time for interactive CAD will be reduced. It is noted that in an earlier study we found no increase in reading times with interactive CAD\(^18\).

To ensure that we obtained enough data for LROC analysis, radiologists were asked to report more findings than they would normally do in screening practice. This might have changed the behavior of the readers compared to routine practice. However, for analysis we used the partial area under the LROC curve for low false-positive fractions. Therefore, reader findings with very low suspiciousness scores did not influence the results.

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. 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 unaided reading, 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.

Acknowledgements

The authors gratefully acknowledge the participation of C.N.A. Frotscher, E. Ghazi, S. Gommers, M.W. Imhof-Tas and U.C. Lalji in the observer performance study.

Bibliography


Computer aided detection of masses: interactive decision support versus prompting


Dual stage presentation of computer-aided detection marks in mammography to reduce both perceptual and interpretation errors

R. Hupse, M. Samulski, R.D. Mus, G.J. den Heeten, C. Boetes and N. Karssemeijer

Original title: Dual stage presentation of computer-aided detection marks in mammography to reduce both perceptual and interpretation errors

Presented at: the Medical Image Perception Conference XIV, Dublin
Abstract

Objective: Recent studies show that the interactive use of CAD, in which prompts are only displayed for queried regions, might be more effective than the conventional use of CAD. In this way interpretation errors are reduced instead of errors due to perceptual oversight. We investigated the combination of traditional and interactive CAD to avoid both type of errors.

Methods: A set of 200 cases (80 cases with a malignant mass) was read by 12 observers in two sessions: without CAD and with interactive CAD. Observers were asked to mark abnormal regions, to assign a malignancy score to each marked region, and to indicate if they would refer the case or not. After completing the interactive session, cases with one or more regions that were marked as highly suspicious by CAD but were not queried or reported by the reader were shown again. Findings and scores for these cases could be changed, based on the CAD prompts. For each observer we computed the mean true-positive fraction (MTPF) in the false-positive fraction interval 0-0.2 of the localization receiver operating characteristic (LROC) curve.

Results: Compared to interactive CAD, dual stage presentation of CAD marks yielded some new cancer detections (18 in total). However, a relatively large number of new false-positives (61 in total) were also added in the second stage. When using interactive CAD the MTPF increased from 0.57 to 0.62 (p=0.01). No additional increase (MTPF of 0.62) was found by reading the subset with potential oversight errors again.

Conclusion: Our conclusion is that for the detection of masses and architectural distortions CAD prompts presented after the interactive session do not improve reader performance.
6.1 Introduction

Many radiologists think that computer-aided detection (CAD) shows too many false-positive prompts to be effective for the detection of masses and architectural distortions. We showed in chapter 4 that the interactive use of CAD, in which prompts are only displayed for queried regions, might be more effective. In this way interpretation errors are reduced instead of errors due to perceptual oversight. In chapter 5 we compared the use of interactive CAD to traditional prompts and found that interactive CAD yielded a significant better performance than the prompting system. Because the use of interactive CAD will only reduce interpretation errors and will not reduce perceptual oversight errors, we investigated if a combination of these systems can improve the results.

The idea is to extend the developed interactive CAD system with a second stage. In the first stage, mammograms are read with the use of interactive CAD. In the second stage, non-referred mammograms with one or more highly suspicious CAD regions are shown again, and the reader can change his or her decision based on the CAD prompts. Using this dual stage presentation, we aim at reducing both interpretation and perceptual errors.

6.2 Methods

6.2.1 Dual stage presentation

The dual stage presentation was included in the observer study described in chapter 5. In the observer study cases were read without CAD and in a mode in which CAD results could be queried interactively. When a reader completed the interactive reading session, a set of mammograms was selected to be shown a second time to the reader. These were mammograms with one or more highly suspicious CAD regions. Only mammograms were selected that were not referred by the reader, and which had highly suspicious CAD regions that were not queried or marked. To select regions, a threshold on the CAD suspiciousness score was set in such a way that on average 0.6 CAD regions were considered highly suspicious in a non-cancer case. The threshold was determined using a large dataset of digitized film mammograms without cancer, not used in the observer study. On data from a previous study\(^1\) (see chapter 4) the selected threshold yielded a selection of 10 percent of the mammograms, which seemed reasonable.

When the selected mammograms were displayed in the second stage, the findings that were marked in the first stage were displayed, as well as prompts for the highly
suspicious CAD regions. Highly suspicious CAD regions that were not queried or marked in the first stage were displayed with a thick contour. If multiple highly suspicious CAD regions were present in the same mammogram, the regions that were already queried or marked were displayed with a thin contour. The interactive aspect of CAD was still enabled in this second stage. After inspection of the CAD prompts, the reader was able to add findings, remove findings and/or change the score for a finding. An example of a cancer that was missed in the first stage but detected in the second stage is shown in figures 6.1a and 6.1b.

6.2.2 Statistical analysis

We counted the number of cancers detected by each reader in the interactive mode before and after the second stage. Differences in this number were analyzed using the Wilcoxon signed-rank test. The number of normal cases with one or more findings was also counted before and after the second stage. Further, we performed a location receiver operating characteristic (LROC) analysis. For each reader we computed the LROC curve using the findings before and after the second stage. LROC curves were computed as described in chapter 5. For each LROC curve we computed a performance measure defined as the mean TPF for FPF’s lower than 0.2. The reason we chose this interval is as follows. In screening settings in the Netherlands, the recall rate is less than 2 percent. However, we used a challenging dataset in which almost 20% of the non-cancer cases were false-positively referred in the original screening.

Differences in the mean TPF were analyzed using the DBM-MRMC method\(^2\). This method treats both readers and data as random samples without taking into account the location of findings. We consider treatment differences with a p-value smaller than 0.05 to be significant only if the global test on the null hypothesis of equal treatments is also significant (p-value < 0.05).

6.3 Results

Figure 7.2 shows for each reader the number of cancer and non-cancer cases that were displayed in the second stage. For most readers, this number is larger than expected. We expected that on average approximately 20 cases would be selected (10% of 200), based on the results of our previous study (see section 6.2.1). Besides the total number of cancer cases, figure 7.2 also shows the number of additional cancers detected by the readers. These are the cases in which the cancer was detected in the second stage while missed in the first stage. For the non-cancer cases the number of cases with a new false-positive finding is shown. Although the second stage yielded some new
6.3 Results

cancer detections (18 in total), a relatively large number of new false-positives (61 in total) were also marked.

Figure 6.3 displays the LROC curves for the unaided mode, interactive CAD and mode with dual stage presentation. These curves are averaged over all readers. In table6.1 the performances for each reader in each mode are listed. DBM-MRMC analysis yielded no significant difference (p=0.87) between the dual stage mode and interactive CAD. A significant difference was found between interactive CAD and unaided reading (p=0.010) and between dual stage mode and unaided reading (p=0.007). The p-value for the global test was 0.010.

Figure 6.1: (a) Craniocaudal images of the right and left breast. In the image of the right breast, an activated CAD region is shown. The malignancy score of the region (65) is given next to the contour. The arrow indicates a malignant mass in the left breast that was missed. (b) In the second stage the case is shown again with 2 very suspicious CAD prompts that were not queried or reported. One of these prompts is located at malignant mass location.

Because the number of cases that were displayed in the second stage was larger than expected, we investigated the effect of reducing this number retrospectively. This was done by setting the threshold on the CAD suspiciousness score in such a way that on average 0.3 instead of 0.6 CAD regions were considered highly suspicious in a non-cancer case. This new threshold yielded a smaller set of selected cases. When repeating the analysis for the second stage, changes made by he readers in the second stage were omitted for cases that were not selected using the new threshold. Figure 6.4 shows the number of selected cases, the number of additional detected cancers and the number of cases with a new false-positive finding for this analysis. When using the lower threshold, only 3 new cancers were detected while 30 new false-positives were
Figure 6.2: The number of cancer and non-cancer cases displayed in the second stage for each reader. For the cancer cases, the number of new detections is shown. For the non-cancer cases, the number of cases with a new false-positive finding is shown.

Figure 6.3: Location receiver operating characteristic (LROC) curves for unaided reading, reading with interactive CAD and for reading with the dual stage presentation. The curves are averaged over all readers.
Table 6.1: Mean true-positive fraction (MTPF) in the false-positive fraction interval 0-0.2

<table>
<thead>
<tr>
<th>Reader</th>
<th>Unaided</th>
<th>Interactive</th>
<th>Dual stage</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.44</td>
<td>0.55</td>
<td>0.55</td>
</tr>
<tr>
<td>2</td>
<td>0.46</td>
<td>0.63</td>
<td>0.65</td>
</tr>
<tr>
<td>3</td>
<td>0.52</td>
<td>0.56</td>
<td>0.57</td>
</tr>
<tr>
<td>4</td>
<td>0.53</td>
<td>0.59</td>
<td>0.60</td>
</tr>
<tr>
<td>5</td>
<td>0.55</td>
<td>0.65</td>
<td>0.65</td>
</tr>
<tr>
<td>6</td>
<td>0.55</td>
<td>0.66</td>
<td>0.66</td>
</tr>
<tr>
<td>7</td>
<td>0.55</td>
<td>0.56</td>
<td>0.56</td>
</tr>
<tr>
<td>8</td>
<td>0.56</td>
<td>0.60</td>
<td>0.60</td>
</tr>
<tr>
<td>9</td>
<td>0.61</td>
<td>0.60</td>
<td>0.59</td>
</tr>
<tr>
<td>10</td>
<td>0.62</td>
<td>0.63</td>
<td>0.63</td>
</tr>
<tr>
<td>11</td>
<td>0.70</td>
<td>0.68</td>
<td>0.68</td>
</tr>
<tr>
<td>12</td>
<td>0.75</td>
<td>0.71</td>
<td>0.71</td>
</tr>
<tr>
<td>Average</td>
<td>0.57</td>
<td>0.62</td>
<td>0.62</td>
</tr>
</tbody>
</table>

marked. Compared to using the original threshold, no positive effect on the LROC curve was observed.

6.4 Discussion

We presented a method for dual stage presentation of CAD marks in mammography that was developed to reduce both perceptual and interpretation errors. While the first stage (interactive support for queried regions) yielded a significant effect on reader performance, no additional effect was found when adding the second stage (display of prompts for missed regions). By presenting CAD prompts for suspicious regions that were not queried or marked, a few more cancers were detected. However, the number of false-positive findings also increased. Therefore, the overall performance of the readers did not increase. Results confirm earlier studies showing errors due to perceptual oversight are rare and are hard to correct by prompts.\(^3-5\).

The number of cases that were selected for display during the second stage was higher than expected. Selection was based on a threshold that was chosen in a way that 10% of the cases from a previous study\(^1\) (see chapter 4) were selected. It was not possible to compute the number of selected cases before starting the study, because selection depends on the user interaction during the experiment. There are multiple differences between this study and the previous study that might affect the number of selected
cases. In the previous study, the majority of the cases (95 out of 120) had only MLO views available. In this study, all mammograms had additional CC views. This yielded more images per mammogram, and therefore more CAD findings per case. Due to this difference, the dataset contained more cases with a CAD score above the threshold. Further, in this study digital mammograms were used while in the earlier study film mammograms were used. We found that CAD scores were, in general, slightly higher for digital images than for film images. We acknowledge that it is not desirable that a large number of cases is presented to the reader in the second stage. This considerably increases reading time and might demotivate readers. The large number of cases that were displayed might have reduced the effect of the second stage because the selection included relatively a larger number of non-cancer cases. However, retrospectively reducing the number of selected cases by using a different selection threshold yielded no positive effect on reader performance. Further research is necessary to investigate the prospective effect of a lower threshold for selecting the cases presented in the second stage.

**Figure 6.4:** Results for simulating a reduction of the number of cases for which the second stage responses were included in the analysis. As in figure 7.2, the number of cancer and non-cancer cases displayed in the second stage, the number of new detected cancers and the number of cases with a new false-positive finding is shown for each reader.
Bibliography


Computer aided detection of masses in mammography: interactive decision support versus independent combination of reader and computer scores

R. Hupse and N. Karssemeijer

Original title: Computer aided detection of masses in mammography: interactive decision support versus independent combination of reader and computer scores

To be submitted
Interactive decision support versus independent combination of reader and CAD

Abstract

Objective: We developed an interactive CAD system for reading mammograms that acts as decision aid. The general effect of interactive CAD seems promising (chapter 5), and on average radiologists’ performance improves with the system. In the past, it was found that independent combination of reader and CAD scores also improves mammogram interpretation. In this study we compare the interactive use of CAD to independent combination of CAD and reader scores.

Methods: We used the locations and scores for regions marked by 12 readers in 200 full field digital mammograms (FFDM) in an observer study. Results from two modes were used: unaided reading and reading with interactive CAD. A weighted sum of CAD and reader scores was used to compute the combined performance. For each reader the mean true-positive fraction (MTPF) was computed for false-positive fractions lower than 0.2 on the location receiver operating characteristic (LROC) curve. Differences in MTPF were analyzed using the DBM-MRMC method.

Results: For three of the 12 readers the performance decreased when interactive CAD was used compared to unaided reading. However, when scores were combined independently, the performance increased for all readers. Nine readers obtained a higher performance with independent combination compared to interactive CAD. No significant difference between independent combination and interactive CAD was found.

Conclusion: The good results of independent combination of reader and CAD scores suggest that most readers should be able to benefit more from CAD than they did in the observer study.
7.1 Introduction

As described in chapter 4 and 5, the effect of computer-aided detection (CAD) techniques that are currently used in many screening programs is not conclusive\textsuperscript{1–7}. Existing CAD systems are developed to mark suspicious lesions in order to reduce search errors. There are indications that for masses the interpretation of a region is more problematic than its detection\textsuperscript{8–10}. Therefore, we developed a new CAD system that acts as decision aid (chapter 4). In chapter 5 we investigated the use of this system in an observer study. We found that there is a large variation in the size of the effect between radiologists. For some radiologists the use of interactive CAD yielded a large increase in performance, while for others a smaller increase or even a decrease was observed. This could be explained by three reasons. First, intra-reader variability can cause differences in performance when reading the same set of mammograms in the unaided and interactive session. This variability can be attributed to environmental factors and to differences in the order in which mammograms are presented. Second, there is a difference in experience between the radiologists. For radiologists that already obtain a relatively high performance for unaided reading there is less room for improvement than for radiologists who perform less well. Third, radiologists might use the interactive CAD system in different ways. This idea is supported by the fact that we observed a large variation in the number of queried regions between the readers.

We expect that radiologists benefit more from interactive CAD when they obtain more experience with the system and learn to use CAD effectively. However, it is hard to predict if the system will be helpful for all radiologists, including the more experienced. One way to investigate the expected benefit is to exclude intra-reader variability and the learning effect by combining reader and CAD scores independently. A previous study by Karssemeijer et al.\textsuperscript{8} shows that independent combination of radiologist and CAD scores yields an improved performance compared to the single reader performance. This method uses a weighted sum of CAD scores and reader scores. We expect that a reader who has learned to use CAD effectively should be able to obtain at least the same performance as obtained by independent combination.

In this study we compare independent combination of reader and CAD scores to the interactive use of CAD for the twelve readers in the observer study described in chapter 5. We hypothesize that the interactive use of CAD yields higher performances than independent combination because readers can incorporate the presence and scores of CAD marks in their own interpretation in an intelligent manner. When a reader obtains a higher performance with independent combination than when using CAD interactively, this suggests that the reader does not make optimal use of CAD information yet. In that case it is expected that the effect of interactive CAD increases when the
reader gets more experience with the system. To optimize the independent combination of reader and CAD scores, we investigate the use of different CAD-weights for experienced readers and for less experienced readers.

7.2 Methods

7.2.1 Observer study

We used the locations and scores for regions marked by 12 readers in 200 full field digital mammograms (FFDM). The selection of cases and the performed observer study are described in full detail in chapter 5. Nine readers were certified screening radiologists and three were residents trained in breast imaging. For each case readers were asked to mark one or more suspicious lesions in the mammogram. A score was given to each lesion on a scale between 0 (not suspicious) and 100 (very suspicious). If a lesion was visible in both views of the mammogram, both locations were marked. The observer study included a mode in which the cases were read without CAD and a mode in which CAD results could be queried interactively. When a region was queried, the contour and suspiciousness score of the CAD system were displayed. If no CAD region was detected at the queried location, nothing was displayed.

7.2.2 CAD system

The CAD system was trained on a large set of digitized film mammograms (11793 images containing 1853 malignant mass regions) as described in chapter 5 and chapter 8. The system was used to detect malignant masses in the study dataset. Each candidate mass region in the study set was assigned a standardized normality score. This score was based on a separate dataset consisting of 1546 images for normal cases. The normal images in this set were not used for training the CAD system and were not used in the observer study. The standardized normality score for a candidate mass region in the study dataset was defined as the number of detected false-positive (FP) CAD regions in the normal set that were equal or more suspiciousness than the region of interest, divided by the number of images in the normal dataset. This yielded a score that is low (for example 0.1 FP/image) for suspicious regions and high (for example 5 FP/image) for less suspicious regions.
7.2 Methods

7.2.3 Independent combination of reader and CAD scores

For each reader, the scores given in the unaided reading mode were independently combined with the CAD scores. This was done by converting the standardized normality score of the CAD region \( \bar{N}_{CAD} \) into a CAD suspiciousness score \( S_{CAD} \):

\[
S_{CAD} = -\log(\bar{N}_{CAD}) \tag{7.1}
\]

and subsequently computing a weighted sum of \( S_{CAD} \) and the reader score \( S_R \):

\[
S_{COM} = S_R + W \cdot S_{CAD} \tag{7.2}
\]

with \( S_{COM} \) the combined score and \( W \) the CAD-weight.

A combined score was computed for each finding marked by the reader. Regions that were detected by CAD but not marked by the reader were not used. If a region was marked by the reader but no CAD region was available for that region, a normality score of 2 FP/image was used for \( \bar{N}_{CAD} \) in equation 7.1. This value was chosen because we wanted to compare results to reading with interactive CAD. In the observer experiment with interactive CAD, the CAD regions with a normality score higher than 2 FP/image were not displayed. Thus, if a reader queried a regions without CAD response, this meant that the score of this region, if CAD had detected it, was 2 FP/image or higher. Therefore, if a CAD region was available with a normality score higher than 2 FP/image, the score was set to 2 FP/image for independent combination. As in the interactive reading session, a CAD-region was considered available if the distance between the CAD region and the marked region of the reader was less than 1.5 cm. If multiple CAD regions were available (for example when a CAD region was present at the marked locations in both craniocaudal and mediolateral oblique views), the most suspiciousness region (lowest normality score) was used.

7.2.4 Adjusting the CAD-weight

The CAD-weight \( W \) used in equation 7.2 was determined using the leave-one-reader-out method, in which the weight for a reader was optimized using the data of all other readers. We compared two different ways of applying the leave-one-reader-out method. First, we considered the 12 readers as a single group and used the data for the 11 other readers when optimizing the CAD-weight for a reader. Second, we splitted the 12 readers into two subgroups of equal size, based on unaided reading performance in the observer study. One subgroup contained the 6 best performing readers, the other subgroup contained the 6 other readers. The performance measure was computed as the mean sensitivity at high specificity values, and is described in more detail in the
next paragraph. When performing the leave-one-reader-out method, $W$ was determined for each reader using the data from the other 5 or 11 readers in the group. This was done by computing the average performance of these readers for a range of values of $W$. The value was chosen for which the average performance was maximal.

It is noted that the way readers used the scale for scoring findings is subjective. This yields relative differences in the optimal-CAD weights for single readers, and might therefore have a negative effect on the leave-one-reader out procedure. To avoid this effect, we repeated the procedure of adjusting the CAD-weights after normalizing the reader scores. In the study by Karssemeijer\(^8\), normalizing was done by dividing the suspicion rating assigned to each of a readers findings by the arithmetic mean of the 25 highest suspicion ratings of the noncancerous findings. Because the number of non-cancer cases in our dataset was lower compared to the referred study (120 instead of 250), we used the mean of the 12 highest ratings for normalizing.

### 7.2.5 Statistical analysis

A performance measure was computed for each reader in three modes: unaided reading, reading with interactive CAD and independent combination of reader and CAD scores. The performance measure was based on the location receiver operating characteristic (LROC) curve constructed using the combined scores $S_{COM}$ (for independent combination) or the reader scores $S_R$ (for unaided reading and reading with interactive CAD). The LROC curve plots the true-positive fraction (TPF) versus the false-positive fraction (FPF). The TPF was computed as the fraction of cancer cases detected. A malignant mass was considered detected if the distance between the marked region and the location of the referral standard was less than 2 cm. This threshold was slightly higher than the threshold of 1.5 cm that was used for determining if a reader finding was also detected by CAD. The reason for this difference is that, for most mammographic images, more CAD regions were present than malignant masses. Therefore, the probability that a reader finding was at the same location as a CAD finding just by chance, was higher than the probability that it was located at a true lesion by chance. If the malignant region was visible in both views (craniocaudal and mediolateral oblique), it was considered detected if the region was marked in at least one of the views. The FPF was computed over all cases without cancer using the most suspicious finding per case. For each LROC curve we computed a performance measure defined as the mean TPF (MTPF) for FPF’s lower than 0.2. This low false-positive range was chosen to match operating points of radiologists in screening. In screening in the Netherlands the recall rate is less than 2 percent. However, in the dataset we used 16.7% of the normal cases (20 of a total of 120) were recalled by the original screening radiologists.
First we compared the two methods described for adjusting the CAD-weights. We decided to use the method in which the readers were divided into two groups only if this would give significant better results. Next, we compared the three modes: unaided reading, reading with interactive CAD and independent combination. Differences in MTPF were analyzed using the DBM-MRMC method\textsuperscript{11}. This method treats both readers and cases as random samples without taking into account the location of findings. We considered treatment differences with a p-value smaller than 0.05 to be significant only if the global test on the null hypothesis of equal treatments is also significant (p-value < 0.05).

### 7.3 Results

Figure 7.1 shows the average reader performance computed in the leave-one-reader-out procedure. Results are shown for the method in which we splitted the 12 readers into two subgroups. Average performances are plotted for each fold of the two groups and for a range of values of $W$.

![Figure 7.1](image-url)

**Figure 7.1:** The average reader performance in the leave-one-reader-out procedure. Results are shown for the method in which we splitted the 12 readers into two subgroups. Average performances are plotted for each fold of the two groups and for a range of values of $W$. Results are plotted for the group of readers with lowest performance (a) for the group of best performing readers (b).

The CAD-weights selected for each reader using the leave-one-reader-out procedure are listed in table 7.1. Weights are given for two methods: method 1 considers all readers as being part of a single group, method 2 considers two groups as described in section 7.2.4. The readers are sorted by unaided performance. For the 6 readers with
highest performance, similar or smaller CAD-weights were found when method 2 was used compared to method 1. For the 6 other readers, larger CAD-weights were found for method 2. This is expected because readers with lower performance are expected to benefit more from CAD-results and therefore need larger weights to obtain best performance for independent combination. The performance measures computed for each reader using the selected CAD-weight are also listed in table 7.1. The presented results are based on the original reader scores. Results obtained using normalized scores are not presented. The average difference in MTPF values for using the original and normalized scores was smaller than 0.005 for both method 1 and 2. For none of the readers a higher performance was found when method 2 was compared to method 1. We decided to use the simplest method, method 1, for comparing independent combination to interactive CAD.

### Table 7.1: CAD-weights optimized using the two leave-one-reader-out methods and mean true-positive fraction (MTPF) for observer study (unaided mode) and independent combination

<table>
<thead>
<tr>
<th>Reader</th>
<th>mean MTPF for observer experiment</th>
<th>CAD-weight, method 1</th>
<th>CAD-weight, method 2</th>
<th>MTPF for method 1</th>
<th>MTPF for method 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lowest performance</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.44</td>
<td>9.7</td>
<td>10.3</td>
<td>0.58</td>
<td>0.58</td>
</tr>
<tr>
<td>2</td>
<td>0.46</td>
<td>9.7</td>
<td>11.6</td>
<td>0.53</td>
<td>0.52</td>
</tr>
<tr>
<td>3</td>
<td>0.52</td>
<td>9.7</td>
<td>9.8</td>
<td>0.62</td>
<td>0.62</td>
</tr>
<tr>
<td>4</td>
<td>0.53</td>
<td>9.7</td>
<td>10.3</td>
<td>0.61</td>
<td>0.61</td>
</tr>
<tr>
<td>5</td>
<td>0.55</td>
<td>9.7</td>
<td>10.3</td>
<td>0.60</td>
<td>0.60</td>
</tr>
<tr>
<td>6</td>
<td>0.55</td>
<td>9.7</td>
<td>10.9</td>
<td>0.62</td>
<td>0.61</td>
</tr>
<tr>
<td>Highest performance</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>0.55</td>
<td>9.7</td>
<td>9.2</td>
<td>0.61</td>
<td>0.61</td>
</tr>
<tr>
<td>8</td>
<td>0.56</td>
<td>9.6</td>
<td>9.0</td>
<td>0.65</td>
<td>0.64</td>
</tr>
<tr>
<td>9</td>
<td>0.61</td>
<td>9.7</td>
<td>9.7</td>
<td>0.64</td>
<td>0.64</td>
</tr>
<tr>
<td>10</td>
<td>0.62</td>
<td>9.7</td>
<td>7.8</td>
<td>0.69</td>
<td>0.68</td>
</tr>
<tr>
<td>11</td>
<td>0.70</td>
<td>9.7</td>
<td>9.7</td>
<td>0.72</td>
<td>0.72</td>
</tr>
<tr>
<td>12</td>
<td>0.75</td>
<td>9.7</td>
<td>9.7</td>
<td>0.77</td>
<td>0.77</td>
</tr>
</tbody>
</table>

Figure 7.2 presents the performance obtained by each reader for the three modes:
unaided reading, reading with interactive CAD and independent combination. As already discussed in the introduction, there is a large variation in the effect of interactive CAD compared to unaided reading. For three of the 12 readers the performance decreased when interactive CAD was used. These three readers were part of the group of best readers. However, when scores were combined independently, the performance increased for all readers. For nine readers (including the 6 best readers) the performance with independent combination was larger than with interactive CAD. The other 3 readers obtained best performance with interactive CAD. Statistical analysis yielded a significant difference between the three modes (p=0.012 for the global test) with a significant difference between interactive CAD and unaided reading (p=0.022) and a significant difference between independent combination and unaided reading (p=0.006). No significant difference between independent combination and interactive CAD was found.

7.4 Discussion and conclusion

Our results show a significant increase in performance when reader scores were independently combined with CAD scores. The increase was found for all readers, even for the readers that already obtained a high performance with unaided reading. For most readers (9 out of 12), independent combination yielded a higher performance than when CAD results were used interactively. This may be due to the fact that readers were inexperienced in using interactive CAD. If this is the case, we expect that the effect of interactive CAD will become larger when readers gain more experience with the system. However, a larger study that includes more cases is necessary to investigate the significance of performance differences for individual readers.

Although we expect that readers should be able to obtain at least the same performance with interactive CAD as with independent combination, it is unknown if the interactive use of CAD will be superior to independent combination. Overall, no significant difference was found between independent combination and interactive use of CAD scores. For 3 out of 12 readers we found a larger performance when using CAD interactively. It may be that these readers found a way to use the CAD results in a very effective manner. It is remarked that intra-reader variability between the unaided reading session and the session with interactive CAD also causes performance differences for independent combination and interactive CAD. A study with more readers will have a greater statistical power for detecting performance differences between the interactive CAD and independent combination.

We found no performance increase by adjusting the CAD-weight for readers with
different levels of performance, compared to using all readers for adjusting the CAD-weight. This is remarkable, because we expected to find that readers with a relatively low performance would benefit more from a larger CAD-weight than from a smaller CAD-weight. For readers with a relatively high performance we expected vice versa. Although slightly higher CAD-weights were selected when the leave-one-reader-out method included only readers with low performance compared to all readers, no performance increase was found when using these higher CAD-weights. The relative small differences between the CAD-weights selected by both methods suggest that CAD is able to improve reader performance in a way that is similar for all readers.

When combining reader and CAD results, regions that were detected by CAD but not marked by the reader were not used. These regions were ignored because of the comparison to interactive CAD, in which CAD results remain hidden until activated by the reader. Therefore, it was not possible to detect cancers that were missed by the reader while detected by CAD. If the CAD system would operate as an individual reader in screening, all CAD findings could be used in the analysis, including the findings missed by the reader due to perceptual oversight. Further research is necessary to investigate this approach.

The good results of independent combination of reader and CAD scores suggest that most readers should be able to benefit more from CAD than they did in the ob-
BIBLIOGRAPHY

server study.

Bibliography


Interactive decision support versus independent combination of reader and CAD

Standalone computer-aided detection compared to radiologists’ performance for the detection of mammographic masses


Original title: Standalone computer-aided detection compared to radiologists’ performance for the detection of mammographic masses

Published in: European Radiology 2012 (in press), DOI: 10.1007/s00330-012-2562-7
Abstract

Objectives: We developed a computer-aided detection (CAD) system aimed at decision support for detection of malignant masses and architectural distortions in mammograms. The effect of this system on radiologists’ performance depends strongly on its standalone performance. The purpose of this study was to compare the standalone performance of this CAD system to that of radiologists.

Methods: In a retrospective study nine certified screening radiologists and three residents read 200 digital screening mammograms without the use of CAD. Performances of the individual readers and of CAD were computed as the true-positive fraction (TPF) at a false-positive fraction of 0.05 and 0.2. Differences were analyzed using an independent one-sample t-test.

Results: At a false-positive fraction of 0.05, the performance of CAD (TPF=0.49) was not significantly different from that of the certified screening radiologists (TPF=0.52, P=0.17). At a false-positive fraction of 0.2, CAD performance (TPF=0.62) was significantly lower than the radiologist performance (TPF=0.74, P<0.001). Compared to the residents, CAD performance was similar for all false-positive fractions.

Conclusion: The sensitivity of CAD at a high specificity was comparable to that of human readers. These results show potential for CAD to be used as an independent reader in breast cancer screening.
8.1 Introduction

To detect breast cancer in an early stage screening programs have been introduced, in which woman are invited to have mammograms taken on a regular basis. The reading of these screening mammograms is highly challenging and requires skilled radiologists. To assist radiologists with this task, computer-aided detection (CAD) systems have been developed and are widely used. In the United States the majority of mammograms are nowadays read with CAD\(^1\). Current CAD systems have been developed to mark regions suspicious for the presence of microcalcification clusters or masses, in order to avoid perceptual oversight of abnormalities by the radiologists. To achieve this goal most systems operate at a high sensitivity. Because the performance of CAD algorithms is still limited, this as has a consequence that the specificity of CAD is relatively low.

In general, radiologists are positive about the use of CAD for the detection of microcalcifications. However, for the detection of masses radiologists have less confidence in CAD, because current systems still show a relatively large number of false-positives\(^2,3\). This might be a reason why some prospective studies do not show significant improvement in reader performance when CAD is used\(^4–7\). It may also be a problem that current CAD systems are only targeting the problem of perceptual oversights. It has been suggested that the misinterpretation of suspicious regions is a more common cause of missed malignant masses\(^8,9\). Because of these reasons, we investigate alternative ways of using CAD, which aim at helping radiologists with interpretation of suspicious regions instead of exclusively focusing on avoiding oversight errors\(^10\), or to use CAD as independent reader\(^11\).

It is evident that the potential benefit of CAD as decision support for the detection of malignant masses depends strongly on the quality of the CAD system that is used. In order to be used for decision support or as independent reader in screening, it would be of great importance if CAD could operate at a sensitivity and specificity comparable to that of a human reader. To achieve this, we are developing a CAD system that has good performance at low false positive rates. In this system novel features are included representing normal tissue context\(^12\) and similarity between mediolateral oblique (MLO) and cranial-caudal view (CC)\(^13\). Furthermore, the system is trained using a very large numbers of normal and abnormal mammograms.

The goal of this study is to explore the quality of our newly developed CAD system by comparing its performance to that of radiologists for the detection of masses. In this study we use the term masses for the group of malignant masses, architectural distortions and asymmetric densities. Data were analyzed obtained in a retrospective study in which 9 radiologists and 3 residents read 200 mammograms acquired in a
Standalone CAD compared to radiologists’ performance

digital screening pilot project\textsuperscript{14}. We compared the standalone performance of CAD to that of the readers at a high specificity comparable to the level used in screening practice.

\section*{8.2 Material and methods}

\subsection*{8.2.1 Computer-aided detection algorithm}

We developed a two-view CAD system as described in\textsuperscript{12,13}. Full field digital mammograms (FFDM) used in this study were not used for training of the CAD system, thus avoiding potential bias. In Figure 8.1 the CAD system is schematically depicted. The first part of the system (including classifier 1 and 2) is a single-view detection stage, the second part (including classifier 3 and 4) is a two-view detection stage. Comparison with prior mammograms is not used by the CAD system in this study. In the single-view stage the breast area and pectoral muscle are segmented and the images are preprocessed. A set of 5 texture features is computed for each location on a regular grid in the image. Based on these location features, a neural network classifier (classifier 1) computes a likelihood score for each location in the image. The likelihood scores together form a likelihood image, and for each local maximum in this image a candidate region is segmented. Subsequently, for each of the segmented candidate mass regions a set of 30 single-view region features is extracted. These features measure region contrast, location, linear texture, density, region size, compactness and normal tissue context. Normal tissue context features are based on the likelihood scores assigned to other areas in the mammogram that are assumed to depict normal tissue\textsuperscript{12}. When these scores are relatively large compared to the scores at the candidate mass location, it is more likely that the candidate mass is a false positive. Using the region features, a neural network classifier (classifier 2) computes a likelihood score for the region.

In the two-view stage each region is compared to all candidate mass regions within a certain search area in the ipsilateral view\textsuperscript{13}. For each pair of regions a set of similarity features is computed. The similarity features are based on the relative location of the regions in the mammogram, correlation between the pixel values, difference in entropy, differences in single-view region features and the output of classifier 2. A 4-class k-nearest neighbour (k-NN) classifier (classifier 3) is used to discriminate four categories of region pairs: true-positive/true-positive, true-positive/false-positive, false-positive/true-positive or false-positive/false-positive. The correspondence score is defined as the likelihood that a region combination represents two true positives. A candidate region is finally linked to the region in the ipsilateral view with the high-
8.2 Material and methods

Figure 8.1: Schematic diagram of the two-view CAD system for mass detection.
est correspondence score, conditioned that this score exceeds a threshold. Then, for
each region a new set of features is constructed combining single-view and similarity features. These features are fed into a fourth classifier which was trained using a case-based learning scheme\textsuperscript{13}.

The described CAD system can be used to detect malignant masses in digitized film images or FFDM images. FFDM images are first normalized using a dedicated processing module\textsuperscript{15} before computing the CAD results. The system was trained on a large set of digitized film mammograms (11793 images containing 1853 malignant mass regions).

8.2.2 Case selection

In the Netherlands, all women between 50 and 75 years of age are invited to participate every two years as part of a free nationwide breast cancer screening service. In the screening, all mammograms are read independently by two radiologists, with referral based on consensus. Details of the program are described elsewhere\textsuperscript{16,17}. The introduction of digital mammography in the screening program was completed in 2010. In this study we used a total of 200 FFDM cases acquired in a digital screening pilot project conducted in the period 2003-2008 at the Preventicon screening centre in Utrecht, the Netherlands\textsuperscript{14}. All material was anonymized and institutional review board approval was waived. As common in the Netherlands, MLO and CC views are obtained at the initial screening. At the subsequent screenings only MLO views are made, unless there is an indication that obtaining the second view would be beneficial.

Mammograms with abnormalities were annotated under supervision of an experienced radiologist 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 a set of 200 cases was selected. The set consisted of 80 cases with a biopsy-proven malignant mass and 120 cases without a cancer. 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 (subtlety score of 1) and cases with only microcalcifications were excluded. First we checked for digital screening mammograms acquired prior to detection in which the mass was already visible (with a minimal sign or more clear). This yielded a total of 17 mammograms. In this paper we will call these cases the cancers that were missed in screening. To obtain a total of 80 cancer cases we added 63 mammograms that were randomly selected from all remaining mammograms in which a malignant mass was detected in an incident screening round. To make the set more challeng-
ing we included 20 non-cancer cases that were falsely referred for further assessment. These 20 cases were randomly selected from the mammograms in which the radiologist reported a suspicious mass but no malignancy was found, further assessment did not include biopsies, and at least one negative follow-up screening mammogram was obtained. Obvious benign abnormalities were not included. The remaining 100 non-cancer cases were digital mammograms that were not referred in the pilot project and had at least one normal follow-up exam. These mammograms were randomly selected while ensuring that the proportion of initial screenings was the same for cases with cancer (4 out of 80) and cases without cancer (6 out of 120).

8.2.3 Study design

All 200 cases were read by 12 readers, of which nine were certified screening radiologists with experience in mammography ranging from 1 to 24 years and three were residents. For analysis of the CAD results, we processed each FFDM case using the two-view CAD system we developed. In the actual observer study, each reader read all cases with and without CAD support. However, for this study we only used reader results obtained when reading without CAD. Before starting the study the readers were briefly trained with a short training session to become familiarized with the system and its functionalities like zooming and contrast enhancement. 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.

Mammograms were processed by the Hologic Selenia FFDM system using standard clinical settings. Processed images were displayed using a 30-inch DICOM calibrated LCD panel (model FlexScan SX3031W; Eizo Nanao Technologies Inc., Hakui, Ishikawa, Japan) with a native resolution of 2,560x1,600. When reading a mammogram, the mammogram made in the screening round previous to the selected screening round (if available) was also displayed. These prior mammograms were only displayed for comparison and could be digital or scanned film mammograms. No findings had to be reported in the prior mammogram. We asked radiologists to report all findings that they considered as potential abnormalities, also those that they would normally not refer in screening practice. This was done to ensure that we could extend our analysis beyond the operating point used in screening. Readers were instructed that an average of about one finding per mammogram would be a good response. A finding was reported by placing a mass icon on the suspicious location. Each finding was numbered. If a finding was visible in both views, radiologists were asked to report the location of the finding in both views. Readers assigned a suspiciousness score in the range 0-100 to each finding. For each case the readers indicated if they would refer the case or not.
8.2.4 Data analysis

Based on the suspiciousness scores given by the readers and by CAD, location receiver operating characteristic (LROC) curves were computed. Sensitivity was computed as the fraction of abnormal cases in which the reader or CAD had reported the mass at the correct location in at least one of the views. The location of a finding was considered correct if the distance to the center of mass 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. In 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 finding with the highest score in each non-cancer case.

To compare the performance of CAD to that of the certified radiologists, we computed the true-positive fraction (TPF) on the LROC curve at false-positive fractions (FPF) of 0.05 and 0.2. Low false-positive fractions were chosen because in screening radiologists work with a high specificity. The TPF values were computed by linear interpolation between the true-positive fractions of the neighboring operating points. The hypothesis that CAD performance differs from the average radiologist performance was tested using an independent one-sample t-test (two-sided).

When CAD is used as independent reader, it should ideally detect cancer cases that are missed by the human readers or identify non-cancer cases that are referred by the human readers. To investigate to what extent the developed CAD system is able to do detect such cases, we also constructed LROC curves for two subsets of the data. The first subset consisted of the cancers missed in screening and all 120 non-cancer cases. The second set consisted of the non-cancer cases referred in screening and all cancer cases.

8.3 Results

The performance of the nine certified radiologists and standalone CAD is shown in Figure 8.2. Results for the three residents and standalone CAD are shown in Figure 8.3. There is a large variance in the sensitivity obtained by the readers. For low false-positive fractions (0-0.05), the TPF of CAD is similar to the mean TPF of the certified radiologists. For larger false-positive fractions (0.14 and higher), all certified radiologists performed better than CAD. Compared to the residents, CAD performance is similar for all false-positive fractions.

Table 8.1 gives the mean, standard deviation and 95% confidence interval of the TPF values obtained by the radiologists at a FPF of 0.05 and 0.2. Statistical analysis indicates that there is no significant performance difference (P=0.169) between standalone CAD
8.3 Results

**Figure 8.2:** Case-based LROC curves for the 9 certified screening radiologists and standalone CAD, computed over the whole dataset of 200 cases.

**Figure 8.3:** Case-based LROC curves for the 3 residents and standalone CAD, computed over the whole dataset of 200 cases.
and the radiologists at a FPF of 0.05. At a FPF of 0.2, the performance of CAD is significant lower (P<0.001) than that of the certified radiologists.

Table 8.1: True-positive fraction (TPF) for the certified radiologists and for CAD, at a false-positive fraction (FPF) of 0.05 and 0.2

<table>
<thead>
<tr>
<th></th>
<th>FPF = 0.05</th>
<th>FPF = 0.2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>TPF radiologists</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mean</td>
<td>0.52</td>
<td>0.74</td>
</tr>
<tr>
<td>standard deviation</td>
<td>0.09</td>
<td>0.06</td>
</tr>
<tr>
<td>95% confidence interval</td>
<td>0.45 - 0.59</td>
<td>0.69 - 0.78</td>
</tr>
<tr>
<td><strong>TPF CAD</strong></td>
<td>0.49</td>
<td>0.62</td>
</tr>
<tr>
<td><strong>p-value</strong></td>
<td>0.169</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

The LROC curves for the subset containing the cancer cases missed in screening and all non-cancer cases are depicted in Figure 8.4. The figure shows the curves averaged over the 9 certified radiologists, averaged over the 3 residents and for standalone CAD. For low false-positive fractions (<0.2), CAD performance is higher than the average performance of the radiologists. An example of a cancer case missed in screening setting is shown in Figure 8.5. Although the subtle mass (finding 1) was detected by 8 certified radiologists and 2 residents, only two readers gave the finding a score higher than 75 and two of the certified radiologists and one resident indicated they would not refer the case. The malignant mass was detected by the CAD system and had been assigned a relatively high suspiciousness score of 82. Findings 2, 3 and 4 are false-positive findings of CAD with scores lower than 50.

Figure 8.6 shows the average LROC curves for the subset consisting of all cancer cases and the non-cancer cases referred in screening. For this subset, an obvious performance difference is found between standalone CAD and the radiologists. For all false-positive fractions less than 0.5, CAD performance is higher than the average performance of the radiologists. An example of a non-cancer case referred in screening is shown in Figure 8.7. In screening the case was referred based on finding 1. This region was marked by 9 readers and 7 indicated they would refer the case. Some certified radiologists assigned very high suspiciousness scores to finding 1 (91 and 80), while CAD assigned a relatively low score of 54 to the same finding. Findings 2 and 3 are false-positive findings of CAD with lower scores.
8.4 Discussion

Results show that the performance of our CAD system for the detection of masses is approaching that of trained radiologists. At a false-positive fraction of 0.05, no significant difference was found between the performance of CAD and the certified radiologists. At a false-positive fraction of 0.2, the radiologists performed significantly better than CAD. However, when standalone CAD was compared to the residents, similar performances were found for all false-positive fractions. To the best of our knowledge, this is the first study in which a CAD system for screening mammograms demonstrated a performance as good as the performance of human readers at a high specificity.

For most cases (190 out of 200) prior screening mammograms were available for the readers, so that they could judge if findings were already visible in a previous screening round. This presentation is similar as in screening practice and can be valuable to detect growing lesions, which are more likely to be malignant. One should take into account that the CAD system we developed did not use any information from prior mammograms. We expect that CAD results will improve when temporal features are included. Another way to improve the CAD system is to use the presence of microcalcifications, as it is highly suspicious when these occur in combination with a mass.

In the observer study, mammograms were displayed on a 4-megapixel color display. This display has less spatial and gray value resolution than the 5-megapixel grayscale monitors used in clinical practice. In principle, this may have affected the

![Figure 8.4: Case-based LROC curves obtained for the subset consisting of the cancer cases missed in screening and all non-cancer cases. The curves are averaged over the 9 certified radiologists and the 3 residents. The LROC for the CAD results is also shown.](image)
Figure 8.5: The MLO (top) and CC (bottom) images for one of the cancer cases missed in screening. Findings of CAD are indicated with circles, findings of radiologists are indicated with squares. Only findings are depicted that had been assigned a score of 40 or more. Finding 1 represents the malignant mass.
8.4 Discussion

Figure 8.6: Case-based LROC curves obtained for the subset consisting of all cancer cases and the non-cancer cases falsely referred in screening. The curves are averaged over the 9 certified radiologists and the 3 residents. The LROC for the CAD results is also shown.

performance of the readers. However, in this study only masses were to be detected and cases with only microcalcifications were excluded. A study from Kamitani et al.\textsuperscript{19} 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.

In a retrospective study readers might perform differently compared to normal screening\textsuperscript{20}. Reasons might be the knowledge that decisions do not affect patient care and there is a competition element to perform better than colleagues. Further, in the original observer experiment cases were read sequentially without and with CAD prompts in the same reading session. When comparing radiologists performance to standalone CAD, we used the reader findings obtained when reading without CAD. It is not certain if decisions for the unaided reading mode reflect decisions that would have been made in an unaided screening setting. It might be that in the observer experiment radiologists relied on the CAD prompts and were less vigilant in the unaided mode. This would introduce a negative bias for the radiologists performance. On the other hand, radiologists might be more alert when they know they are part of a study, which would introduce a positive bias for the radiologists performance. In our study we found that the average reading time per case (44 seconds) was at least as long as in screening practice. Therefore, we have no indication to believe that readers were
Figure 8.7: The MLO (top) and CC (bottom) images for one of the non-cancer cases referred for further assessment. Findings of CAD are indicated with circles, findings of radiologists are indicated with squares. Only findings are depicted that had been assigned a score of 40 or more. The case was referred based on finding 1.
less vigilant in the unaided mode than they would have been if the same cases were read in screening. Nevertheless, large scale real life prospective studies are necessary to confirm this.

Another limitation of the study is that the presented results are obtained for a limited set of 200 cases. We selected a challenging set of cases for this study. Cases with obvious masses were excluded, as well as microcalcification cases in which no mass or architectural distortion was visible. In our study, the proportions of normal and abnormal cases were different than in screening practice. Therefore we expect that decision thresholds differ from practice. However, we focused on LROC analysis, which is not affected by decision thresholds. Our results indicate that, compared to radiologists, standalone CAD performed better for non-cancer cases that were referred in original screening and cancer-cases that were missed in original screening. This suggests that the performance difference between standalone CAD and radiologists is very dependent on the dataset analyzed. Because the dataset is so different from screening practice, it is hard to translate absolute performance differences found in this study to screening practice. Prospective studies will be necessary to compare readers and CAD performance for larger datasets representing a real life screening population. Besides the non-referred normal cases and the detected cancers, these studies should also include all missed cancer cases and the referred non-cancer cases.

A CAD system that operates at a specificity and sensitivity comparable to that of a radiologist has great potential. For instance, it can be used as an independent reader next to the screening radiologists\textsuperscript{11}. In screening, double reading of mammograms by two radiologists is common practice and is an effective method to improve mammographic interpretation\textsuperscript{21–24}. Improved performance has also been found when interpretations of up to 12 radiologists were independently combined\textsuperscript{25}. This means there is a large potential to improve mammogram reading, even when double reading is already practiced. An application of standalone CAD can be to select a small percentage of mammograms that obtained a high CAD suspiciousness score but were not referred by double reading. These mammograms could be selected for additional reading by a third radiologist. We expect that this will improve cancer detection without a large increase in false-positives because our results show that part of the cancer cases missed in screening were detected at a very low false-positive level. Another possibility is to use CAD for decision support during the reading session. Previous research showed that this improved reader performance\textsuperscript{10}.

In summary, we developed a CAD system for the detection of malignant masses that is aimed at decision support. At a high specificity, no significant difference was found between standalone CAD performance and that of certified radiologists. This suggests this CAD system might be useful as an independent reader in screening.
Acknowledgements

The authors gratefully acknowledge the participation of C.N.A. Frotscher, E. Ghazi, S. Gommers, U.C. Lalji, R.M. Mann and R.D. Mus in the observer performance study.

Bibliography


Standalone CAD compared to radiologists’ performance


Publications

Papers in international journals


Papers in conference proceedings


Other publications


Summary

To detect breast cancers in an early stage, in most western countries screening programs are organized. A small fraction of the mammograms acquired in screening contain malignancies, which can be very subtle. In order to reduce oversight errors, computer-aided detection (CAD) systems have been developed. The effect of CAD techniques that are currently used in screening programs is not conclusive. The research described in this thesis had two goals. The first goal was to improve the performance of current CAD techniques for mass detection. The second goal was to investigate the use of CAD as decision support instead of perception aid.

When reading mammograms, radiologists do not only look at local properties of suspicious regions but also take into account more general contextual information. Chapter 2 described a set of context features that represent suspiciousness of normal tissue in the same case. When suspiciousness scores for normal tissue are relatively large compared to the scores at a candidate mass location, it is less likely that the candidate mass is a malignant tumor. Context features were computed for three normal reference areas defined in the image at hand, in the contralateral image and in different projections (if available). Evaluation showed that the mean sensitivity increased more than 6% when context features were added. Context computed using multiple views yielded a better performance than using a single view. Besides the importance of using multiple views, results showed that best CAD performance was obtained when multiple context features were combined that are based on different reference areas in the mammogram.

Chapter 3 addressed the problem of selecting a useful subset of features from a large number of available region descriptions. Feature selection depends on the choice of a performance measure used as optimization criterion. We compared two optimization criteria, one was computed as the mean sensitivity for low false-positive fractions and reflects the use of CAD in clinical practice. The other criterion was a standard approach using the general test statistic Wilks’ lambda. It was found that significantly higher performances were obtained when feature sets were selected by the standard approach. These results indicate that a general statistic as Wilks’ lambda is more powerful for selecting a good feature set.

Chapter 4 presented a new way of displaying CAD results to the radiologist. Current CAD systems make use of prompts that are intended to reduce perceptual oversight errors. However, in practice, many masses are not missed by perceptual oversight but due to incorrect interpretation. A CAD system was presented in which CAD
marks and their associated suspiciousness scores remain hidden unless their location is queried by the radiologist. To evaluate this approach, an observer study was performed in which four screening radiologists and five non-radiologists read 120 cases of which 40 cases had a malignant mass that was missed at the original screening. The average sensitivity of the readers at low false-positive rates significantly increased from 25.1% to 34.8% when interactive CAD was used. Reading time was not affected.

In chapter 5, the effect of interactive display of CAD results was compared to the effect of conventional prompting. An observer study was performed in which nine screening radiologists and three residents twice read 200 screening cases, once with CAD prompts and once with interactive CAD. In the prompting mode, findings were recorded before and after activating CAD. The study set contained 63 cases with screen-detected malignant masses, 17 cases with a malignant mass missed at the original screening, 20 cases that were falsely referred at the original screening and 100 normal cases. The average sensitivity of the readers at low false-positive rates was used as a measure of reader performance. Averaged over all readers, the performance for unaided reading was 57% and increased significantly with interactive CAD (62%), while remaining unaffected by prompting (57%). These results show that 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 reducing perceptual oversights.

The interactive CAD system described in chapter 4 and 5 has potential to reduce interpretation errors only, because CAD marks remain hidden unless their location is queried by the radiologist. In chapter 6, we investigated a dual stage presentation of CAD marks to reduce both perceptual and interpretation errors. In the first stage, mammograms were read with the use of interactive CAD. In the second stage, non-referred mammograms in which one or more highly suspicious CAD regions occurred that were not queried, were presented again to the reader, with the non-queried regions displayed as prompts. By doing this, the reader could still change his or her decision in case regions marked by CAD were overlooked. The dual stage presentation was included in the observer study described in chapter 5. It was found that, compared to interactive CAD, dual stage presentation of CAD marks yielded some new cancer detections (18 in total). However, a relatively large number of new false-positives (61 in total) were also added in the second stage. No additional increase in Location Receiver Operating Characteristic (LROC) performance was found by reading the subset with potential oversight errors again.

In chapter 7 the interactive use of CAD was compared to independent combination of CAD and reader scores. We hypothesized that the interactive use of CAD yields higher performances than independent combination because readers can incorporate
the presence and scores of CAD marks in their own interpretation in an intelligent manner. For the comparison we used the locations and scores for regions marked by the readers in the observer study described in chapter 5. The combined performance was computed using a weighted sum of CAD scores and reader scores given in the un-aided mode. Compared to unaided reading, independent combination yielded higher performances for all readers. For nine readers (including the 6 best readers) the performance with independent combination was larger than with interactive CAD. No significant difference in performance between independent combination and interactive CAD was found. The good results of independent combination of reader and CAD scores suggest that most readers should be able to benefit more from CAD than they did in the observer study.

The effect of a CAD system aimed at decision support depends strongly on its standalone performance. Chapter 8 compared the standalone performance of the developed CAD system to that of radiologists. Performances of nine certified screening radiologists that took part in the observer study described in chapter 5 were compared to the performance of CAD. This comparison was done at a high specificity comparable to the level used in screening practice. At a false-positive fraction of 0.05, the true-positive fraction of CAD was 0.49 and did not significantly differ from the average true-positive fraction of the certified screening radiologists (0.52, p=0.169). At a false-positive fraction of 0.2, the true-positive fraction of CAD (0.62) was significantly lower than the radiologist performance (0.74, p<0.001). Compared to the residents, CAD performance was similar for all false-positive fractions. These results show potential for CAD to be used as an independent reader in breast cancer screening.
Samenvatting

Om borstkanker in een vroeg stadium op te sporen, worden in de meeste westerse landen screeningsprogramma’s georganiseerd. Een klein deel van de mammogrammen die in screening gemaakt worden bevatten kwaadaardige afwijkingen. Deze afwijkingen kunnen heel subtiel zijn. Om perceptiefouten in de detectie van borstkanker te voorkomen, zijn computer-aided detection (CAD) technieken ontwikkeld. Het effect van de CAD-technieken die momenteel gebruikt worden in screening is echter niet overtuigend. Het onderzoek dat beschreven staat in dit proefschrift had twee doelen. Het eerste doel was om de detectie van tumorschaduwen door CAD te verbeteren. Het tweede doel was om te onderzoeken of CAD gebruikt kan worden om fouten bij de interpretatie van verdachte gebieden te verminderen, in plaats van het verminderen van perceptiefouten.

Bij het lezen van mammogrammen kijken radiologen niet alleen naar lokale eigenschappen van verdachte regio’s, maar houden ook rekening met meer algemene contextuele informatie. Hoofdstuk 2 beschrijft een set van context features die de verdachtheid van normaal weefsel in dezelfde patiënt weergeven. Wanneer de verdachtheid van normaal weefsel relatief groot is ten opzichte van een potentiële tumor locatie, is de kans kleiner dat de potentiële tumor locatie een maligniteit bevat. We berekenden context features voor drie referentiegebieden in het beeld zelf, in het contralaterale beeld en in verschillende projecties (indien beschikbaar). We vonden dat de gemiddelde sensitiviteit met meer dan 6% steeg wanneer context features werden toegevoegd. De features die berekend waren met informatie uit meerdere beelden van dezelfde patiënt, gaven een grotere verbetering dan de features die berekend waren met informatie uit een enkel beeld. Verder lieten de resultaten zien dat de grootste verbetering werd verkregen wanneer er meerdere context features tegelijk werden gebruikt die gebaseerd zijn op verschillende referentiegebieden in het mammogram.

Hoofdstuk 3 beschrijft het probleem van het selecteren van een subgroep uit het grote aantal features dat beschikbaar is voor iedere potentiële tumor regio. Deze selectie hangt af van de maat die gekozen wordt als optimalisatie criterium. We hebben twee optimalisatie criteria vergeleken. De eerste was de gemiddelde sensitiviteit voor hoge waarden van specificiteit en weerspiegelt het gebruik van CAD in de klinische praktijk. Het andere criterium was een standaard aanpak met behulp van de algemene test statistiek Wilks’ lambda. De features die geselecteerd werden met behulp van de standaardmethode gaven een significant beter resultaat. Dit geeft aan dat een algemene test statistiek als Wilks’ lambda meer power heeft voor het selecteren van een
Samenvatting

goede feature set.

In hoofdstuk 4 onderzochten we een nieuwe manier van het weergeven van CAD-resultaten aan de radioloog. De CAD-systemen die momenteel gebruikt worden in screening markeren alle verdachte gebieden in een mammogram en richten zich daarbij uitsluitend op het verminderen van perceptiefouten. Perceptiefouten zijn fouten waarbij de radioloog een verdacht gebied over het hoofd heeft gezien. In de praktijk worden veel tumorschaduwen niet gemist door perceptiefouten, maar door een verkeerde interpretatie van een verdacht gebied. We hebben een CAD-systeem ontwikkeld waarin CAD-markers en de bijbehorende waarden van verdachtheid onzichtbaar blijven, totdat de radioloog deze informatie opvraagt door met de computer-muis te klikken op verdachte gebieden. Als er CAD-informatie beschikbaar is op de opgevraagde locatie, wordt deze alleen voor het desbetreffende gebied getoond. Om deze aanpak te evalueren, werd een studie uitgevoerd waarin vier screeningsradiologen en vijf niet-radiologen 120 mammogrammen lasen, waarvan 40 een kwaadaardige tumorschaduw hadden die gemist was tijdens de oorspronkelijke screening. Wanneer het interactieve CAD-systeem werd gebruikt, steg de gemiddelde sensitiviteit van de lezers significante van 25.1% tot 34.8% bij een vast gekozen laag fout-positief doorverwijzingspercentage.

In hoofdstuk 5 werd het effect van interactief CAD vergeleken met het effect van de traditionele manier om CAD-resultaten weer te geven zoals dat momenteel in de klinische praktijk wordt gebruikt. Een studie werd uitgevoerd waarin negen screeningsradiologen en drie niet-radiologen 200 mammogrammen lasen. Dit gebeurde in twee verschillende sessies, een sessie met het traditionele CAD-systeem, en een sessie met interactief CAD. In de sessie met het traditionele CAD-systeem werden de bevindingen van de lezers gevraagd voor en na het activeren van CAD. De gelezen serie bestond uit 63 gevallen met een kwaadaardige tumorschaduw die tijdens de oorspronkelijke screening was gedetecteerd, 17 gevallen met een kwaadaardige tumorschaduw die was gemist in de oorspronkelijke screening, 20 gevallen die ten onrechte waren doorverwezen en 100 normale gevallen. Gemiddeld over alle lezers, steg de sensitiviteit bij een laag fout-positief doorverwijzingspercentage significante van 57% voor het lezen zonder CAD naar 62% voor het lezen met interactief CAD. Het traditionele CAD-systeem had geen effect op de sensitiviteit. Deze resultaten tonen aan dat voor de detectie van kwaadaardige tumorschaduwen, CAD als ondersteuning van de interpretatie van verdachte gebieden effectiever kan zijn dan het huidige gebruik van CAD ter voorkoming van perceptuele fouten.

Het interactieve CAD-systeem dat beschreven werd in hoofdstuk 4 en 5 heeft alleen potentie om interpretatiefouten te verminderen. Het systeem kan geen perceptiefouten verminderen omdat CAD markeringen onzichtbaar blijven totdat de locatie wordt
opgevraagd door de radioloog. In hoofdstuk 6 onderzochten we het effect van een CAD-systeem met twee fases dat zowel perceptiefouten als interpretatiefouten zou kunnen verminderen. In de eerste fase werden mammogrammen gelezen met interactief CAD. In de tweede fase werden mammogrammen die niet waren doorverwezen en waarin één of meer zeer verdachte CAD-gebieden niet waren opgevraagd, opnieuw aan de lezer getoond met daarin de markeringen voor de betreffende zeer verdachte gebieden. Door dit te doen, kon de lezer zijn of haar beslissing nog veranderen in het geval dat de betreffende regio’s over het hoofd waren gezien. Het CAD-systeem met twee fases werd gebruikt in de studie die beschreven werd in hoofdstuk 5. De tweede fase van dit systeem leverde in totaal 18 nieuw gedetecteerde tumorschaduwen op. Echter, er werd ook een relatief groot aantal regio’s door de lezers toegevoegd (61 in totaal) die geen tumorschaduw bleken te bevatten. Hierdoor gaf het opnieuw lezen van de gevallen met potentiële perceptiefouten geen verbetering van de beoordeling.

In hoofdstuk 7 hebben we het interactieve gebruik van CAD vergeleken met de onafhankelijke combinatie van scores gegeven door CAD en door lezers. Onze hypothese was dat het interactieve gebruik van CAD betere resultaten zou geven dan de onafhankelijke combinatie, want lezers kunnen de aanwezigheid en de scores van CAD op een intelligente manier verwerken in hun eigen interpretatie. Voor de vergelijking hebben we de locaties en de scores gebruikt voor de regio’s die gecorreleerd waren door de lezers in de studie van hoofdstuk 5. Voor het onafhankelijk combineren gebruikten we de lezer markeringen van vóórdat de traditionele CAD resulaten waren geactiveerd. In vergelijking met lezen zonder CAD leverde het onafhankelijk combineren een beter resultaat voor alle lezers. Voor negen lezers (inclusief de 6 beste lezers) was het resultaat zelfs beter dan met interactief CAD. We vonden geen significante verschil tussen onafhankelijke combinatie en het interactieve gebruik van CAD. De goede resultaten van de onafhankelijke combinatie van lezer en CAD-scores suggereren dat de meeste lezers meer moeten kunnen profiteren van CAD dan zij deden in de studie.

Het effect van een CAD-systeem als ondersteuning bij de interpretatie hangt sterk af van de prestatie van het systeem zelf. Hoofdstuk 8 vergeleek de prestatie van het ontwikkelde CAD-systeem met de prestatie van radiologen. Hiervoor gebruikten we de markeringen en scores die, zonder de hulp van CAD, door de negen screeningsradiologen waren gegeven in de studie van hoofdstuk 5. De vergelijking werd uitgevoerd bij een hoge specificiteit, vergelijkbaar met de specificiteit in de screeningspraktijk. Bij een fout-positief doorverwijzingspercentage van 5% werd 49% van de abnormale gevallen gedetecteerd door CAD. Dit was niet significant afwijkend van de gemiddelde sensitiviteit van de screeningsradiologen (52%). Bij een fout-positief doorverwijzingspercentage van 20%, werd 62% van de abnormale gevallen gedetecteerd door
Samenvatting

CAD. Dit was significant lager dan de gemiddelde prestatie van de radiologen (74%). De prestatie van CAD was vergelijkbaar met de prestatie van de residents ongeacht de waarde van het fout-positieve doorverwijzingspercentage. Deze resultaten tonen aan dat CAD mogelijk gebruikt zou kunnen worden als onafhankelijk lezer in borstkankerscreening.
Dankwoord

Er zijn veel mensen die mij geholpen hebben bij het uitvoeren en opschrijven van mijn promotieonderzoek. Allereerst natuurlijk mijn promotor, prof. Karssemeijer. Nico, zonder jouw kennis, ervaring, hulp en tijd was dit proefschrift niet geweest wat het nu is. Ik heb de afgelopen jaren veel waardering gehad voor de manier waarop je mij begeleidde. Ons wekelijks overleg was zeer nuttig, en je wist me met jouw ervaring vaak de juiste richting op te sturen. In het begin was onze groep nog relatief klein, en ik herinner me nog goed hoe we met behulp van een whiteboard door jou uitgelegd kregen hoe de software in elkaar stak. Ondertussen is de groep flink gegroeid en ben jij professor. Naast al de extra verplichtingen bleef je toch tijd houden voor ons overleg.

Het KWF-project, van waaruit mijn onderzoek werd gefinancierd, was mede aangevraagd door prof. Boetes. Het plan was dat zij mijn tweede promotor zou worden. Een aantal keer heeft ze deelgenomen aan ons overleg, en haar jarenlange ervaring als radioloog is inspirerend geweest voor de ideeën die in dit proefschrift zijn uitgewerkt. Carla, ik vind het jammer dat je mijn promotie niet meer mee kan maken.

Het was prettig de dagelijkse beslommeringen te kunnen delen met mijn collega’s. Michiel, je achtervolgde me vanuit de stage voor onze masteropleiding naar de afdeling radiologie, en hebt altijd bij mij op de kamer gezeten. Bedankt voor de gezelligheid, je hulp en het aanhoren van mijn geklaag als er eens iets tegenzat. Bedankt dat je mijn paranimf wil zijn! Maurice, we hebben nauw samengewerkt bij het opzetten van de observer experiment en uiteindelijk samen meerdere artikelen gepubliceerd. Het was gezellig om met je naar Groningen te rijden om daar ons systeem te installeren. Ik zal niet snel vergeten hoe we daar met de hele handel in het verkeerde ziekenhuis stonden. Gelukkig konden we er wel om lachen. Bedankt dat je telkens klaar stond met technische hulp! Ik wens je het allerbeste in Noorwegen. Guido, je begon als aio toen ik al even bezig was. Ook jij stond telkens klaar om te helpen. Het was erg prettig om ervaringen te kunnen delen met betrekking tot het ouderschap en solliciteren.

Zonder de vrijwillige medewerking van een aantal mensen waren hoofdstuk 5 tot en met 8 nooit tot stand gekomen. Ard, Carolien, David, Elham, Marc, Mechli, Ritse, Roel, Ruud, Suzanne en Ulrich, bedankt voor jullie inzet tijdens het omvangrijke experiment!

Curriculum Vitae

Rianne Hupse was born 3 June 1981 in Gouda, the Netherlands. She completed secondary school in 2001 and started the study Medical Natural Sciences at the Free University of Amsterdam. After she obtained her B.Sc. degree in 2004, she started the master Cognitive Neuroscience at the Radboud University Nijmegen and obtained her M.Sc. degree in 2006 (cum laude). In September 2006, she started as a PhD student at the Radiology department of the Radboud University Nijmegen Medical Centre. The results of the work she carried out in the Diagnostic Image Analysis Group of that department are described in this thesis. Rianne is married to Bert Swart. Together they have one son: Joël (2010).