Datalogisk Institut
KØBENHAVNS UNIVERSITET


# Mammographic Pattern Recognition

Jakob Raundahl

**Technical Report no. 08-04**
**ISSN: 0107-8283**

**Dept. of Computer Science**
University of Copenhagen • Universitetsparken 1
DK-2100 Copenhagen • Denmark

# Mammographic Pattern Recognition

by

Jakob Raundahl

PhD Dissertation

The Image Group, Department of Computer Science
Faculty of Science, University of Copenhagen

## Preface

This constitutes Jakob Raundahl's PhD thesis. The thesis is submitted October 2007 at the Department of Computer Science, Faculty of Science, University of Copenhagen in partial fulfillment of the requirements for the degree of doctor of philosophy.

The PhD project started April 2003 in a collaboration between the IT University of Copenhagen (ITU) and Center for Clinical and Basic Research (CCBR), each contributing with one third of the funding. The last part was funded by the danish Center for IT research. At the end of 2006 the project moved, with the rest of the former Image Group at ITU, from ITU to the Department of Computer Science at University of Copenhagen. The work has been supervised by Professor Mads Nielsen and Assistant Professor Marco Loog.

The general purpose of the project was to investigate the use of automated image analysis and pattern recognition to help identify patients with increased risk of developing breast cancer. We approached this problem by investigating, building upon, and extending the concept of mammographic density, which has been shown in numerous studies to be related to breast cancer risk.

## Contact Information

The author of this dissertation can currently be contacted at:

Jakob Raundahl
DIKU
Universitetsparken 1
2100 Copenhagen Ø
Denmark
Mobile: +45 20 76 67 27
Email: raundahl@gmail.com

## Acknowledgements

I would like to thank the Center for Clinical and Basic Research, in particular director Claus Christiansen and radiologist Paola Pettersen, for providing the opportunity to make this investigation. My supervisor Mads Nielsen has with his positive outlook and ingenuity contributed throughout the progress of the project. My cosupervisor Marco Loog has managed to continuously provide sharpminded and focused feedback to the endless requests for comments and suggestions. I would also like to thank Nico Karssemeijer from Radboud University Nijmegen Medical Centre, who is an expert in the field of mammographic image analysis and who took the time to come to Copenhagen to hear about the project and give his advice. Furthermore, he showed me great hospitality on my visit to Nijmegen. Finally, I want to thank my colleague Lars Conrad-Hansen without whom the four years would have felt a lot longer.

# Contents

# List of Figures

# List of Tables

# Part I

# Background and Preliminaries

# Chapter 1

# Introduction

Breast cancer is one of the most serious diseases among women in the western world. It is the most common and deadly cancer for women on a global scale, where breast cancer accounts for 21% of all cancer cases and 14% of all cancer deaths [2]. Early detection is critical to the chance of successful treatment, since the cancer may be invasive and spread to the rest of the body [3]. Such detection is typically very difficult, since the first signs of breast cancer are often asymptomatic. This is why x-ray mammography is widely used to screen for breast cancer. The mammogram can show small changes in breast tissue which may indicate cancers which are too small to be detected either by the patient or by a doctor.

The general purpose of this PhD project is to help identify patients who have higher risk of developing breast cancer, based on screening by mammography. Such identification leads to better allocation of screening resources and thereby earlier cancer detections and lower mortality rates. Moreover, the ability to assess relative risk from mammographic patterns would be an important tool as safety measure in clinical trials. We approach this problem by investigating, building upon, and extending the concept of mammographic density, which has been shown in numerous studies to be related to breast cancer risk.

This chapter contains a short overview of the project background and motivation. In addition, an overview of the contents of the thesis is provided.

## 1.1 Some project background

Center for Clinical and Basic Research (CCBR) is a private research institute mainly investigating conditions and diseases that appear in the years after the menopause and has received worldwide recognition for its clinical research especially in the area of osteoporosis and bone disease. CCBR performs research for the pharmaceutical industry and is involved in a number

of the major multi-centre protocols for the development of new drugs for prevention and treatment of osteoporosis and climacteric complaints.

One of these investigations concerns hormone replacement therapy (HRT), which is one of the treatments to help women who have menopausal symptoms. The use of HRT has in recent years become a controversial subject and large, national scale HRT trials have been stopped or cancelled due to studies showing that the hormones had more negative than positive effects [4]. A relative risk of 1.24 for invasive breast cancer (95% CI, 1.01-1.54) was found in the Womens Health Initiative randomized controlled trial [5] for those on combined hormone therapy compared with placebo. HRT also increases mammographic density [6, 7, 8, 9], which has been associated with increased risk of cancer. To what extent these two effects are causally linked, however, is still unclear. It is also an open issue whether all types of HRT increase risk since different kinds of hormone therapy and routes of administration seem to have different influences on the breast density [10].

Mammograms are acquired as a safety measure in HRT trials, mainly investigating osteoporosis, conducted by CCBR. Collecting these data gives a great opportunity to retrospectively analyse mammographic density and its relation to HRT. The ambition is not to automate an existing density measure, but to come up with entirely new measures, so as to capture different aspects of breast changes extending the existing concept of breast density.

## 1.2 Motivation and goals

This project was motivated by the increasing need for automated methods in radiology as a whole and by the challenge of contributing with a new and more specific method to evaluate breast density. There are several advantages of automated methods. The most obvious are that they save time and are often more reproducible. Another potential benefit is the ability to include features of the image that cannot be seen by the radiologist. The interest in breast density arose from its link with increased risk of breast cancer. An automated measure of breast density could be a great advance in this field, where the visual approach using four categories performed by the radiologist is still the final word.

The common ways to evaluate new automated density measures are either through visual assessment or correlation with radiologist readings, which means that even though advanced and powerful image analysis methods are applied the endpoint is still an approximation of a radiologist giving a score of 1-4 based on visual assessment. One of the ways we contribute to the already vast amount of breast density and risk research is by applying a new idea for evaluation of new measures.

Our idea is to develop and evaluate new automated measures directly

based on their ability to separate data expressing difference in density. This enables us to move beyond the existing, crude categorical scores not only providing a continuous rating but also allowing entirely new measures. In this way, the focus has been moved from existing measuring apparatus to the processes under investigation.

Since some types of HRT has been shown to increase mammographic density, images from HRT studies can be used to evaluate density measures by their ability to separate the HRT and placebo populations. The more indicative measure will do the most sensitive and specific classification of HRT and placebo. The route of administration and the combination of hormones will lead to different results and therefore analyses performed by a trained radiologist, blinded to the labels, is eventually needed to validate a new automated method

Based on this reasoning we study images from an HRT trials, building a framework to develop measures indicative of HRT. These measures are then tested as predictors of breast cancer risk on data from a breast cancer study. In the same way, we can potentially learn patterns indicative of breast cancer risk from the breast cancer study and test if these patterns change in different HRT studies.

In our work, initial experiments using a relatively simple thresholding technique were performed establishing that the effect of HRT can indeed be detected based on a small data set of 25 hormone treated patients and 25 control patients. Subsequently more data were collected and a more complex measure based on unsupervised clustering was developed, eventually leading to an elaborate supervised framework.

To get a better understanding of the nature of the mammographic data and the potential of various methods, new methods were applied to more data as soon as promising results were obtained on the initial data set. This caused a simultaneous development of methods and application to more and more data.

## 1.3 Overview of the thesis

After two introductory chapters, the thesis is split up in two major parts; one focusing on the methodological aspects of the research and the other on application to clinical data and discussions derived therefrom. Taking the methodology for granted, the second major part may be read separately from the first. Another way to phrase this division is that the first part details *how* measures are devised and the second part considers the clinical impact of the devised measures. The thesis concludes with a chapter containing an overall summary and discussion. For clarity the introductory chapters are labelled **Part I**, the conclusion **Part IV**, and the two major parts **Part II** and **Part III**.

The thesis contains the following parts and chapters:

**Part I**

1. **Introduction**
   The present chapter introducing the project.

2. **Mammographic Density**
   The technique of mammography is briefly explained followed by a discussion and description of mammographic density, the link to breast cancer risk, and the possibly associated link to hormone replacement therapy. Finally, the chapter contains an overview of different automated approaches at measuring mammographic density and mammographic patterns.

**Part II**

3. **Automated Thresholding Method**
   A basic density measurement is presented and evaluated. The technique is based on global thresholding, basically dividing the breast tissue into two regions, a dense with intensities above the threshold and a non-dense with intensities below. The idea to use thresholding for density assessment was proposed by Byng et al. [11] in 1994 albeit using an interactive method, where an expert sets the threshold manually. The chapter contains work originally published at the SPIE Medical Imaging conference in 2004 [12] and demonstrates a significant increase in density for HRT-users after two years of treatment.

4. **Unsupervised Method**
   In this chapter another automated method for measuring the effect of HRT w.r.t. changes in mammographic patterns of the breast is presented. Unsupervised clustering of features, describing the elongatedness of local image structure, is employed to divide a mammogram into four structurally different areas. Subsequently, based on the relative size of the areas, a density score is determined. Results using the method are presented together with possible interpretations of the measure. Comparisons to the automated threshold and two types of radiologist's readings are included. The chapter contains material published at the SPIE Medical Imaging conferences in 2006 and 2007[13, 14] and at the International Workshop on Digital Mammography in 2006 [15].

5. **Supervised Method**
   We propose a supervised approach and demonstrate that it can be

trained to detect changes due to aging and HRT. Because of the supervised machine learning approach employed, the method can be adapted to the detection of other mammographic changes. The chapter contains material [16] accepted for publication at the MICCAI 2007 conference.

6. **Supervised framework extended using SFS feature selection**
   We present a framework for incorporating feature selection in our supervised methodology. This framework is applied to a set of data from the Dutch national breast cancer screening program. The presented results demonstrate the ability and potential of including feature selection to improve and specialize measures. This work is currently being drafted for publication.

**Part III**

7. **Comparing the effects of orally and nasally dosed HRT on mammographic density and patterns**
   Here we show that pulsatile hormone therapy via the nasal administration route may provide relative advantages in terms of breast safety compared with the apparent adverse effects of oral hormone therapy. Secondarily, it is shown that automated computer-based analysis of digitised mammograms provides a sensitive measure of hormone-induced changes in breast density which could be useful a monitoring tool in future clinical trials assessing the safety of estrogen or hormone replacement therapies. The chapter contains material [17] accepted for publication in Climacteric.

8. **Automatic scoring of mammographic patterns is more indicative of estradiol treatment than breast density**
   We show that estradiol induces changes not only of the mammographic density, but also in the mammographic patterns. These subtle changes are measured in a computerized fashion. Patterns relating to estradiol treatment were more indicative than BI-RADS and percentage density. Percentage density was not significantly more indicative of HRT than BI-RADS, but had significantly lower intra-observer variability. There was no significant difference in patterns shown indicative of breast cancer risk between groups. This work is currently being drafted for publication.

9. **Local pattern scoring of mammograms is a strong and independent predictor of breast cancer**
   In this chapter we demonstrate the following. Percentage density is more indicative of breast cancer risk than BIRADS. Actually, BIRADS proved redundant when adjusted for percentage density. The local

pattern scoring is shown to be more indicative of risk than percentage density. Percentage density and pattern scoring are independent and an aggregate measure combining their separate information effectively doubled the odds ratios of standard density alone. Patterns indicative of HRT was not found to be indicative of risk. This work is currently being drafted for publication.

**Part IV**

10. **Conclusion**
    This chapter contains a general summary of the thesis, a discussion of the findings, and finally a short overall conclusion.

Summarizing, there are three ways to read a meaningful subset of the thesis: parts I, II, and IV for the methodological details; parts I, III, and IV for the clinical details; and parts I and IV for just the big picture.

# Chapter 2

# Mammographic Density

In this chapter the technique of mammography is briefly described. The concept of mammographic density is discussed in more detail, followed by a discussion of the link to breast cancer risk. Finally, an overview of different automated approaches for assessing breast density is given.

## 2.1 Mammography

The primary use of mammography is in the screening and diagnosis of breast cancer. The goal is to detect clinically occult breast cancer at a smaller size and earlier stage than it would otherwise be detected in an effort to interrupt the natural history of breast malignancies and reduce the number of women who die each year from breast cancer.

In the 1960s, the first randomized controlled trial of screening with mammography was initiated in a health insurance program in New York, to test whether screening asymptomatic women for breast cancer could lower the death rate. The trial involved 62,000 women between 40 and 64 years of age. By comparing the subsequent number of deaths among the screened women with those in the control group, the investigators demonstrated that early detection could decrease the mortality from breast cancer [18]. Now, most western countries have national programs to offer annual or bi-annual screenings to women above a certain age.

In mammography each breast is compressed to a thickness of approximately 6 cm and an x-ray is taken perpendicular to the plane of compression. Radiologists generally obtain two projective views corresponding to a horizontal and an oblique vertical compression. These views are called the craniocaudal (CC) and the mediolateral oblique (MLO) respectively. The process is illustrated in Figure 2.1.

The images are analyzed in symmetrically positioned pairs of same projection of left and right breast. This helps the radiologists detect the early

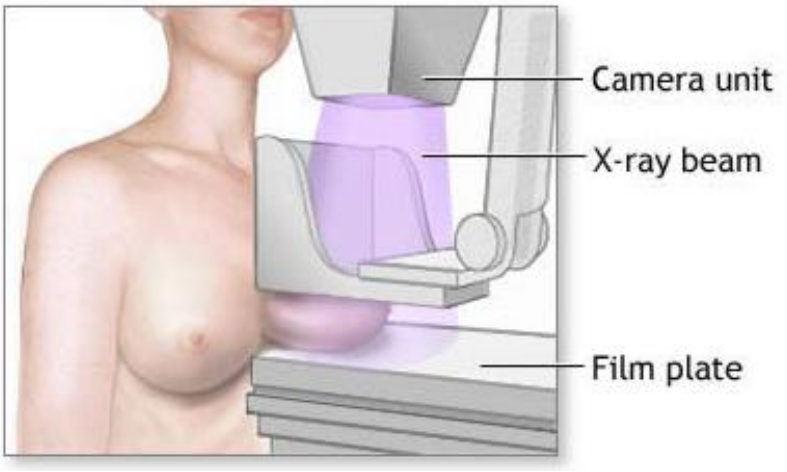



Figure 2.1: The acquisition of a CC mammogram where the projection of the breast tissue is along the vertical axis.

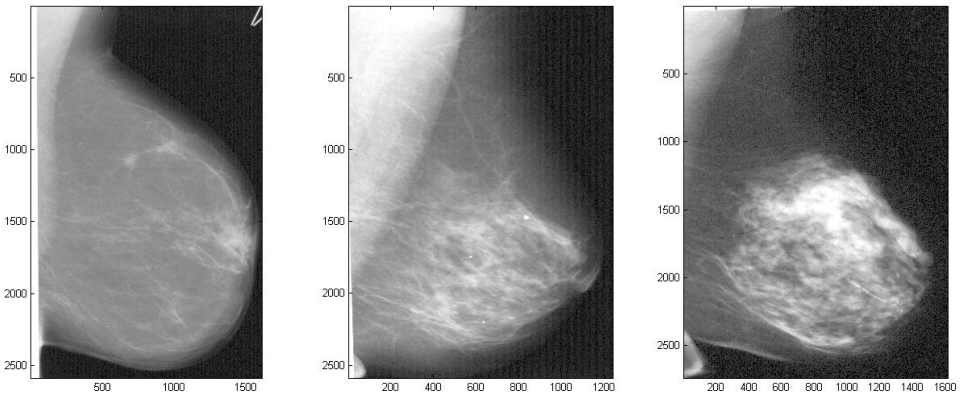
signs of breast cancer. Signs of breast cancer usually shows on mammograms as one or a combination of the following [3]:

- Neodensity or new calcifications
- Mass
- Calcifications: focal or segmental
- Asymmetry: focal or more diffuse asymmetric density
- Architectural distortion

If suspicious findings are present at the screening, a patient will be recalled for a diagnostic session.

## 2.2 Mammographic Density

Mammographic density refers to the prevalence (and to some degree the distribution) of fibroglandular tissue in the breast as it appears on a mammogram. The fibrous and glandular tissues cannot be distinguished in mammography due to a combination of physiological intertwining and similar x-ray attenuation coefficients. These tissues can, however, be distinguished from fatty tissue, which attenuates x-rays to a lower degree. This causes the fibroglandular tissue to stand out as bright areas on a dark background and therefore the term density is used to describe its appearance. Some examples of mammograms with different densities are displayed in Figure 2.2.




(a) (b) (c)

Figure 2.2: Three examples of mammograms with different density
(a) Low density; (b) Medium density; (c) High density

The first to correlate mammographic density with risk of breast cancer was J N Wolfe [19], and his research in the mid 1970's lead to an early, four category, classification scheme now referred to as Wolfe patterns. This method is based on qualitative, visual assessment and contains the following four classes corresponding to ascending magnitude of risk:

**N1:** The breast consists of mostly fatty tissue with no ducts visible. This category represents an essentially normal breast and is considered a lower risk pattern.

**P1:** The breast consists of mostly fatty tissue, but with predominant ducts in the anterior portion covering up to a quarter of the area. It is considered a low risk pattern.

**P2:** The breast is involuted with prominent duct patterns of moderate to severe degree, occupying more than one fourth of the breast area. The visible duct pattern can occupy the entire breast. It is considered a high risk pattern.

**DY:** This category represents a dense parenchyma[1]. It can appear homogeneous due to the overall increased density and the prominent duct pattern cannot be seen. It is considered the highest risk pattern.

In recent years, a similar classification method proposed by The American College of Radiology has become more used. This density classification is a modified version of the Wolfe patterns and is named the BI-RADS

[1]The collection of components that constitute a gland is referred to as the parenchyma

density score [20]. BI-RADS scoring mainly communicates the effect of the dense tissue on the diagnostic sensitivity and discerns the following four categories:

**I:** The breast is almost entirely fat.

**II:** There are scattered fibroglandular densities that could obscure a lesion on mammography.

**III:** The breast is heterogeneously dense. This may lower the sensitivity of mammography.

**IV:** The breast tissue is extremely dense, which lowers the sensitivity of mammography.

This classification is part of a standard mammogram reading in the USA.

In addition to these types of ratings, people have been using planimetric approaches to get a more precise estimate of the amount of dense tissue in the breast. Manual tracing of the boundaries of the dense regions in the mammogram and inter-active thresholding are the two main approaches to count the number of dense pixels in the mammogram . For an in-depth review of these methods and their history the reader is referred to [21].

### 2.2.1 Mammographic density and its relation to breast cancer risk and use of HRT

Numerous studies on the relation between mammographic density, measured with the just described methods, and breast cancer risk have been reported. Boyd et al. [22] have investigated 15 independent studies (ten case-control studies and five cohorts or case-control studies nested in cohorts[2].) with a total of 6,274 patients with breast cancer and 11,638 controls. These studies show that women with high breast density appear to have up to a six fold increase in breast cancer risk corroborating Wolfe's original premise.

Unlike the other known similar or greater breast cancer risk factors[3], breast density can be influenced. HRT has been shown in several studies to increase mammographic density [6, 7, 8, 9]. Therefore an important question is whether or not the relationship between density and breast cancer risk is causal, i.e. a change in the density of a patient's breasts corresponds to a change in breast cancer risk. If the relationship is causal this knowledge could be used in a potential, preventive treatment strategy to lower the density. So far, at least two types of anti-estrogen drugs have been shown

[2]Nested case-control studies draw their cases and controls from cohort populations that have been followed for a period of time.

[3]Genetic abnormalities, age, prior breast cancer, history of breast cancer in first-degree relatives, and biopsy findings [21]

to lower the breast density (tamoxifen [23] and raloxifene [24]). A causal relation would also mean that different types of HRT, and other drugs potentially altering the density, would be easier to classify as safe or unsafe directly based on density findings.

Whether changes in density correspond to changes in cancer risk has been investigated [25] and recent results indicate that this is indeed the case [26]. It is still unclear whether this holds for all types of density changes since findings by Boyd et al. [27] indicate that the effects of hormone therapy on mammographic density, and on breast cancer risk, are separate and not related causally. The reason these, perhaps separate, effects are hard to investigate may be the crudeness of the applied density measures. There are many different ways breast tissue may appear more dense on a mammogram. Both local, focal or multifocal, changes or more diffuse manifestations of dense tissue can lead to identical increases in overall density.

### 2.2.2 Approaches to automated assessment of mammographic density

The Gail model [28] is a popular risk assessment tool and uses a woman's own personal medical history (number of previous breast biopsies and the presence of atypical hyperplasia[4] in any previous breast biopsy specimen), her own reproductive history (age at the start of menstruation and age at the first birth of a child), and the history of breast cancer among her first-degree relatives (mother, sisters, daughters) to estimate her risk of developing invasive breast cancer over specific periods of time. Recent reports using breast density assessed by BI-RADS and continuous, planimetric measures found that the addition of breast density to the Gail model increased its ability to predict cases of breast cancer [29, 30].

Currently, however, the density is not used to assess risk in standard clinical screening procedures or included in general breast cancer risk assessment tools. A reason behind this is that, while breast density has become a well-established risk factor, the best way to measure, and indeed what exactly to measure, is still a debated research topic [31].

The manual, categorical scores have a difficult time struggling both with a low number of categories and large inter- and intra-observer variation [32, 33]. Together with the obvious advantages of automated methods these shortcomings have spurred different branches of density quantification research. The following is a brief overview of these approaches. The discussion of specific, existing literature will be elaborated in the relevant chapters when we present related methods.

The BI-RADS and planimetric measures lead to the most straightfor-

[4]Hyperplasia is a general term referring to extraordinary proliferation of cells within an organ or tissue

ward quantification scheme, which is estimating and counting the bright, dense pixels. There have been reported several automated approaches to this problem ranging from thresholding techniques [34, 35] over phantom- and wedge-based techniques [36] to rigorous modelling based on medical physics [37] and training of statistical models based on extracted features [38, 39, 40]. An extension to this scheme involves estimating the 3D volume of dense tissue in stead of just the area [41]. There has also been an investigation into the individual correlation of different features of the image with radiologists's grading of mammographic density [42]. We present an automated thresholding method quantifying the area of dense tissue in Chapter 3.

Common to the validation of these approaches are the lack of texture information. The original Wolfe classification, however, was partly based on a quantitative property (the amount of dense tissue) and a more qualitative textural property (the appearance and distribution of the ductal pattern). Using texture analysis as basis for automatically classifying breast tissue was originally investigated by Miller and Astley [43] who found it might be viable for detecting and quantifying glandular tissue. Moreover, recent findings have indicated that breast cancer risk is affected not only by the amount of mammographic density but also by the degree of heterogeneity of the breast pattern and, presumably, by other features captured by the Wolfe classification [44]. This has led to a more pattern recognition based approach in which a set of training images are used to build a statistical model based on image features and Wolfe labels [45]. Training on Wolfe labels, however, still means that the endpoint is an approximation of a radiologist giving a score of 1-4 based on visual assessment.

We propose a novel approach in which a measure based in some way on local structure and texture of the image is evaluated directly based on its ability to separate treatment versus non-treatment or high-risk versus low-risk. This means not aiming at reproducing radiologist's scores with the potential advantage of finding new manifestations of treatment and potential risk factors in the mammograms. This way of approaching the problem is presented using an unsupervised framework in Chapter 4, a supervised framework in Chapter 5, and finally supervised framework including feature selection in Chapter 6.

# Part II

# Development of Methods

# Chapter 3

# Automated Thresholding Method

In this chapter a basic, automated method is presented and evaluated with respect to separation of HRT and placebo. Initial results are presented which add more evidence for HRT induced increase in mammographic density. The technique is based on global thresholding, basically dividing the breast tissue into two regions, a dense with intensities above the threshold and a non-dense with intensities below. The density score is then defined as the ratio between the dense area and the total area of breast tissue. This measure is sometimes called the percentage density and has been shown to be a better discriminator of future risk with greater reproducibility than categorical scores [46].

There are typically two projective viewpoints available of the breast in mammography. The cranio-caudal view (CC), where the breast is compressed horizontally and an x-ray is taken in the direction from head to toe, and the medio-lateral (ML), where the breast is vertically compressed and x-ray is taken from the side. An earlier study have indicated, that either projective view can be used when measuring the mammographic density [47], but is this also true for detecting, perhaps small, temporal changes? This is also investigated in these initial experiments.

The idea to use thresholding for density assessment was proposed by Byng et al. [11] in 1994 albeit using an interactive method, where an expert sets the threshold using slider and a screen showing the dense area corresponding to current slider position. More recently, two automatic threshold algorithms have been presented. Zhou et al. [34] employ a method in which the histogram of the image is classified into one of four classes. Subsequently, a global threshold is calculated based on this classification. Sivaramakrishna et al. [35] propose an automatic method, based on Kittler and Illingworths optimal threshold [48], for estimating breast density.

Section 3.1 describes the thresholding methodologies applied in study.

Section 3.2 presents the data used in the experiments. Section 3.3 presents the experimental setup and demonstrates that a proposed heuristic method outperforms methods based on the Kittler and Illingworth's threshold. Additionally, this section shows that the average density measured with the mean-based threshold method increased significantly from 1999 to 2001 in the HRT group ($p < 0.001$). Finally, Section 3.4 discuss the findings and conclude that the ML and CC projections proved equally good for separating HRT and Placebo while using the images of the left breast was better (but not significantly) for separation than using the right.

## 3.1 Methodology

The aim of this method is to isolate the bright, dense areas on the mammograms and calculate their combined area. The measure should be robust and simple providing a quick estimate of the density variations in the investigated data.

Due to intensity variations in the mammograms, caused by varying individual exposure times and compression thicknesses, a fixed threshold for all images performs poorly. We will not use the histogram modelling presented by Zhou et al., since it is not desirable that the measure is very sensitive to the appearance of the histogram, changing algorithm if a borderline unimodal histogram becomes multimodal or vice versa, when investigating longitudinal data. Instead we implemented Kittler and Illingworth's optimal threshold. It was applied to the variance normalized image as suggested by Sivaramakrishna et al. and, in addition, the performance of applying the thresholding algorithm directly on the segmented breast tissue region was investigated.

Another ad hoc approach was developed in which the threshold value is calculated as a factor times the average intensity of the breast. The actual factor (same factor used for all cases) was found in a heuristic approach. Pixels with an intensity higher than this threshold are classified as representing dense tissue. The procedure is as follows:

1. Delineate the breast tissue in the mammogram.
2. Determine the threshold, $T$, as 1.3 times the average intensity.
3. Threshold the image and compute the area of pixels with value higher than $T$.
4. Return the density as this area divided by the total breast area.

The average intensity used to compute the threshold varies with the density of the breast. This causes a gradual underestimation of the actual dense area. Nevertheless, this primitive measure works surprisingly well.

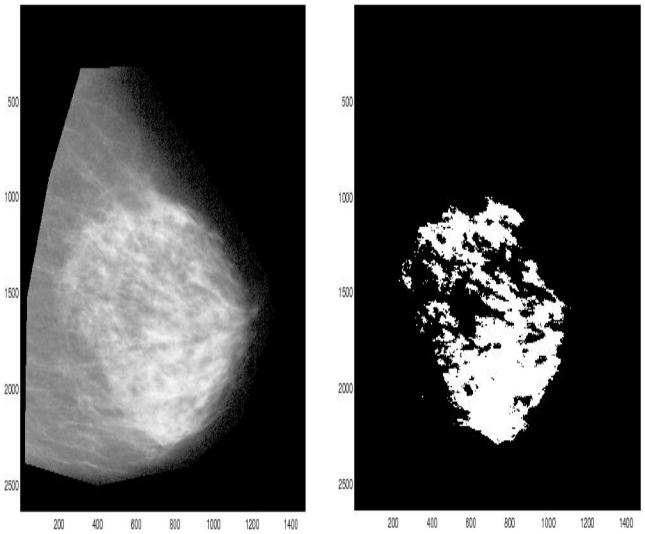



Figure 3.1: Illustration of segmentation by automatic thresholding
(a) The input image; (b) The segmented dense tissue shown in white

## 3.2 Materials

Mammograms from 50 women enrolled in a 2-year prospective, randomized, double-blind, placebo controlled HRT study [49], were digitized for these initial experiments. Double-blind means that both the study subjects and those who interact with the subjects are unaware of whether the subject is in the treatment group or the control group. The participants received either HRT ($N = 25$) or placebo ($N = 25$). Mammograms were acquired at baseline in 1999 and at the end of the study in 2001. The 50 women were randomly selected among the study completers having both baseline and follow-up mammogram.

Breast images were acquired using a Planmed Sophie mammography X-ray unit. The images were then scanned using a Vidar scanner to a resolution of approximately 200 microns with 12 bit gray-scales. Delineation of the breast boundary on the digitized image was done manually using 10 points along the boundary connected with straight lines, resulting in a decagon region of interest.

Since HRT has been shown to increase mammographic density [6, 7, 8, 9] these images can be used to evaluate density measures by their ability to separate the HRT and placebo populations. The methods for measuring density will differ only in which combination of mammographic viewpoints[1] that are used, and in that way an evaluation of the amount of

[1]Here viewpoints refer to both the ML and CC projective views and the views of the left and right breast

density change information, that is carried in each view, is performed.

## 3.3 Experimental setup and results

We want to investigate which view or combination of views give the best separation of placebo and HRT patients under the assumption: the larger the change in density, the higher the probability of an HRT patient. The feature used for separation is the change in density from baseline in 1999 to the end of the study in 2001,

$$\Delta D = D_{2001} - D_{1999}$$

If one view is used for separation, a patient is represented by one feature and if two views are used, two features characterize each patient. To construct the ROC curves, the likelihood of being an HRT patient are calculated from the feature(s).

We base our estimation of the class conditional probabilities of the observations on a simple Gaussian model assuming different class covariances, which corresponds to well-known quadratic discriminant analysis [50].

The ability of a certain combination of views to separate the HRT group from the placebo group is evaluated with areas under ROC curves (AUC). ROC stands for "Receiver Operating Characteristic" and ROC curves depict the performance of a diagnostic test. In short, they show the sensitivity of the test as a function of the specificity. A test which is useless for making a diagnosis will, on average, yield a straight line from (0,0) to (1,1). A useful test will curve close to the upper left corner of the graph. The areas under ROC curves can be compared to determine the relative utility of different tests.

The standard deviations of the AUCs are estimated using a bootstrap scheme [51]. In each iteration 25 samples from the HRT group and 25 samples from the placebo group are chosen at random with replacement. Then an ROC diagram is computed using the randomly chosen data sets, the area is stored and the algorithm moves on to next iteration. This is done until the mean and variance of the areas converge, thereby simulating having enough experiments to estimate the standard deviation of the AUCs. 1000 iterations proved enough for convergence in the experiments. Two AUCs are considered significantly different if there is no overlap of their means $\pm$ one standard deviation.

The benefit from having both projective views for a single breast is compared to that of having an image for both the left and the right breast using identical projections. Combining the left and right view is an approximation of the statistical benefit of having two independent measurements, since the left and right breast are considered symmetric organs when doing

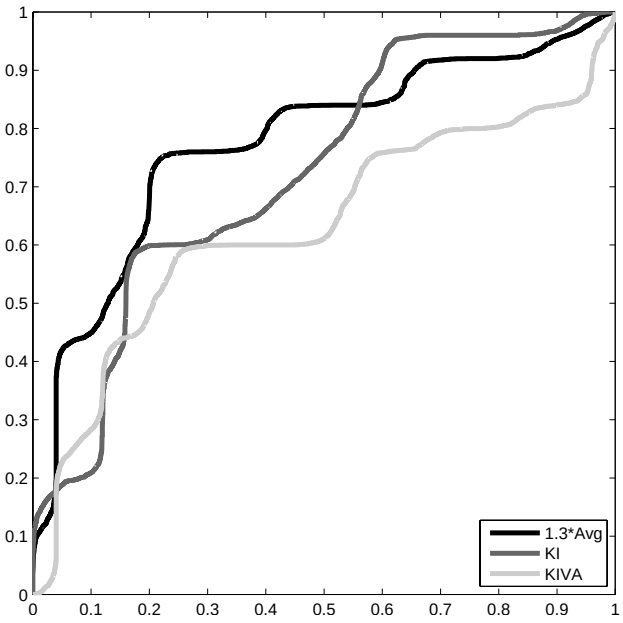



Figure 3.2: ROC curves for classifying HRT and placebo based on the measured change in density in the HRT group.

mammographic analysis [3]. If the separation using two projective views is significantly better than using the same view of left and right breast we will conclude that there is carried independent information in each projection regarding the temporal changes of the mammographic density.

### 3.3.1 Comparison of thresholding algorithms

In order to present the results as clearly as possible only the best performing of the three thresholding algorithms is used in the viewpoint experiments. The three algorithms are the Kittler and Illingworth's optimal threshold (KI), KI applied to the variance normalized image as suggested by Sivaramakrishna et al. (KIVA), and the adaptive threshold based on the mean breast intensity (1.3*Avg). An experiment using the left ML projections indicated that the performance of the mean-based method was superior to that of the other two methods. ROC curves for classifying HRT and placebo based on the measured change in density in the HRT group are shown in Figure 3.2. The AUCs are 0.78, 0.73, and 0.63 for the 1.3*Avg, KI, and KIVA thresholdings respectively. Based on this the mean-based method was used in the viewpoint experiments.

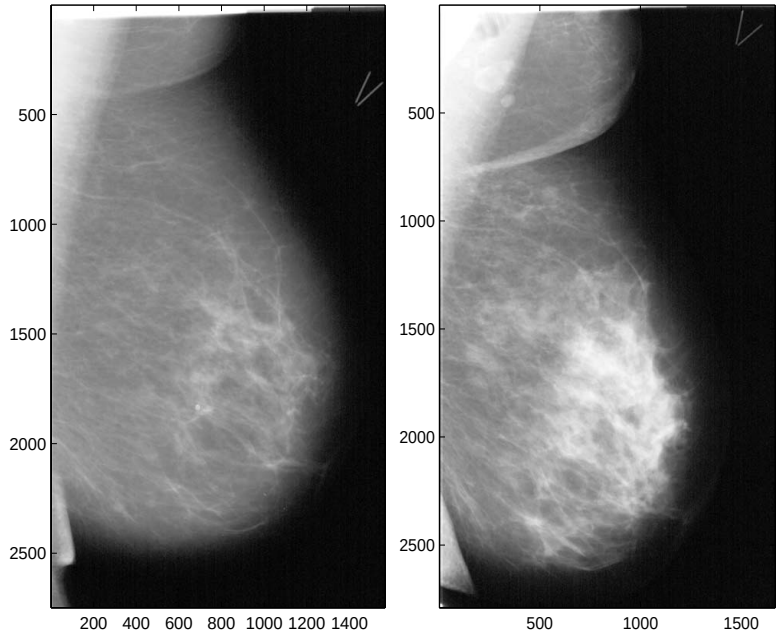



Figure 3.3: Two mammograms from a patient receiving HRT. The picture on the left is acquired in 1999, and the one on the right is of the same breast two years later.

### 3.3.2 Projective viewpoint and changes in density

The results of the density measures can be seen in table 3.1. The HRT density averages have increased significantly from 1999 to 2001 ($p < 0.001$), whereas the change in the placebo group is not statistically significant ($p > 0.1$). A standard, paired t-test was used to test for statistical significance. The increase in density is illustrated by an example in Figure 3.3.

Table 3.2 shows the AUC's for all views and combinations of two views. The separation using two views is better than when using one for 7 out the 8 viewpoints. The overlapping standard deviations indicate that more samples are needed to show this with strict statistical significance. The improvement from using a combination of the two projective views, ML and CC, is similar to the statistical benefit of using a left and right version of *one* perspective view (on average the ROC area increases from $0.76 \pm 0.04$ to $0.79 \pm 0.04$ in both cases). The changes in density are shown as scatterplots in Figures 3.4 and 3.5. A visual inspection of these also suggests that the separation of the placebo and HRT groups are about the same in the two cases. Using more than two views did not increase the performance significantly.

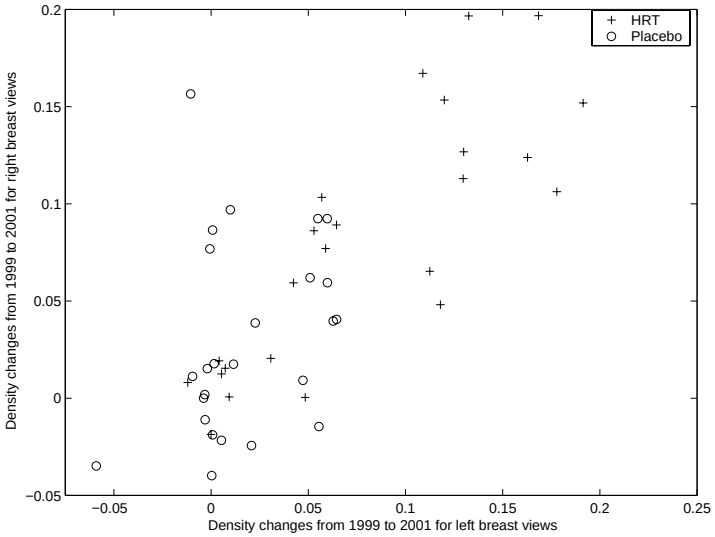



Figure 3.4: Scatter-plot of the density changes of left breasts (mean change of ML and CC) versus the changes of right breasts.

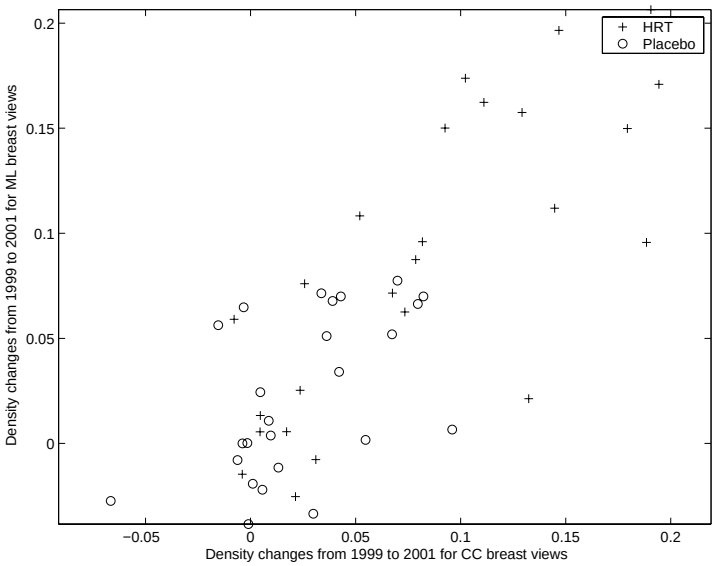



Figure 3.5: Scatter-plot of the density changes of CC projections (mean change of left and right) versus the changes of ML projections.

Table 3.1: Measured percentage density means and standard deviations of the means (STDOM). The view is to be read as ”[Year][(L)eft/(R)ight breast][Projective view]”.

| View | Placebo mean | Placebo STODM | HRT mean | HRT STDOM |
|---|---|---|---|---|
| 99LML | 8.6 | 1.4 | 7.5 | 1.7 |
| 01LML | 11 | 1.9 | 17 | 2.4 |
| 99LCC | 9.8 | 1.4 | 6.9 | 1.6 |
| 01LCC | 11 | 1.6 | 15 | 2.0 |
| 99RML | 12 | 1.4 | 8.9 | 1.4 |
| 01RML | 14 | 2.1 | 17 | 2.1 |
| 99RCC | 10 | 1.7 | 8.0 | 1.6 |
| 01RCC | 14 | 2.1 | 17 | 2.2 |

Table 3.2: Measured AUC’s and bootstrapped standard deviations. The areas have been multiplied with 100.

| Viewpoints | View 1 | View 2 | Both |
|---|---|---|---|
| LML and LCC | $78 \pm 6$ | $81 \pm 6$ | $82 \pm 6$ |
| RML and RCC | $74 \pm 7$ | $72 \pm 7$ | $75 \pm 7$ |
| LML and RML | $78 \pm 6$ | $74 \pm 7$ | $79 \pm 6$ |
| LCC and RCC | $81 \pm 6$ | $72 \pm 7$ | $79 \pm 6$ |

## 3.4 Discussion and conclusion

These initial results have added more evidence for HRT induced mammographic density increase. After two years of treatment the average density of the HRT population was significantly higher than that of the placebo group ($p < 0.001$). A slight, but not significant, increase in density was observed in the placebo group.

The results show that the benefit of having two views compared to one is the same whether it is two projective views or using both left and right breast (ROC areas increase with 0.03). This suggests some symmetry in the evolution of dense tissue. Perhaps asymmetric findings in the evolution

of density in the ML and CC views can be good markers of abnormalities that need attention. Much in the same way as radiologists already look for asymmetry in mammograms of left and right breast [3]. This knowledge could be a useful addition to an automated mammography screening feedback unit.

We think the poor performance of the KIVA approach is due to unstable variance images used in the normalization step. Using KI alone, although performing better than KIVA, generally segmented very large areas of the breast. This might be because the KI algorithm is designed for more general foreground versus background segmentation.

The method based on the mean intensity is quite simple compared to existing attempts at measuring the density, and more efforts could be made to further tweak performance and robustness. On the other hand, planimetric approaches are very developed and we think that more indicative measures and better understanding of effects are possible using more complex methodologies including structural and textural information. Furthermore, there are findings indicating that breast cancer risk is affected not only by the amount of mammographic density but also by the degree of heterogeneity of the parenchymal pattern and, presumably, by other features captured by the Wolfe classification [44]. Therefore we proceed in the following chapters to develop measures able to capture structure.

# Chapter 4

# Unsupervised Method

Using automated thresholding we have demonstrated that HRT induced density changes in our data. We think more indicative measures and better understanding of effects are possible using more complex methodologies including structural and textural information. In this chapter we proceed to develop a pattern recognition based density measure able to quantify the changes caused by HRT in a more indicative way.

The developed measure is still evaluated directly based on its ability to separate treatment versus non-treatment, and we are not aiming at reproducing radiologist's scores. This has the potential advantage of finding new manifestations of treatment and potential risk factors in the mammograms. The resulting measure may differ from existing conceptions of mammographic density and therefore we will use the term mammographic pattern relaxing the strict correspondence to tissues of bright intensities.

This leaves us with a very open problem and many potential approaches. One approach, presented in this chapter, is the use of unsupervised clustering to investigate whether an intrinsic subdivision of the breast tissue may be used as a mammographic pattern score discriminating HRT and placebo patients. The reason behind this unsupervised clustering approach is that the data properties are unknown to us. In this situation, unsupervised clustering is a standard approach of 'letting the data organize itself' [50].

Other groups have applied unsupervised clustering as an intermediate step to compute mammographic density [52, 45]. This work mainly differs in methodology by the choice of features and the way several clustered areas are combined to directly form a mammographic pattern score.

In Section 4.1 the overall methodology is described and we conclude by establishing that features with some invariance properties outperforms features focusing on a visually pleasing clustering. In Section 4.2 the measure based on the clustering of invariant features is compared to the automatic thresholding method and to two types of radiologist's readings, BI-RADS and interactive percentage density. The performance is better than for both

automatic thresholding and BI-RADS and comparable to that of percentage density. Finally a discussion of the findings is presented in Section 4.3.

## 4.1 Methodology

To generate a representative collection of features we sample features from random positions in every image in the data set. Subsequently, these features are clustered producing $N$ classes. Based on these $N$ labels a classifier is trained to label all pixels, potentially from new images, resulting in images with pixel values from one to $N$. The density of a given image is then computed based on each of the $N$ areas in the image corresponding to the $N$ classes.

The overall methodology can be summed up like this:

- Extract features representing the entire data set
- Apply clustering to label the extracted features
- Classify all pixels in each mammogram based on the outcome of the clustering
- Determine relative areas of the classes for each mammogram
- Determine a general scoring as a combination of these areas

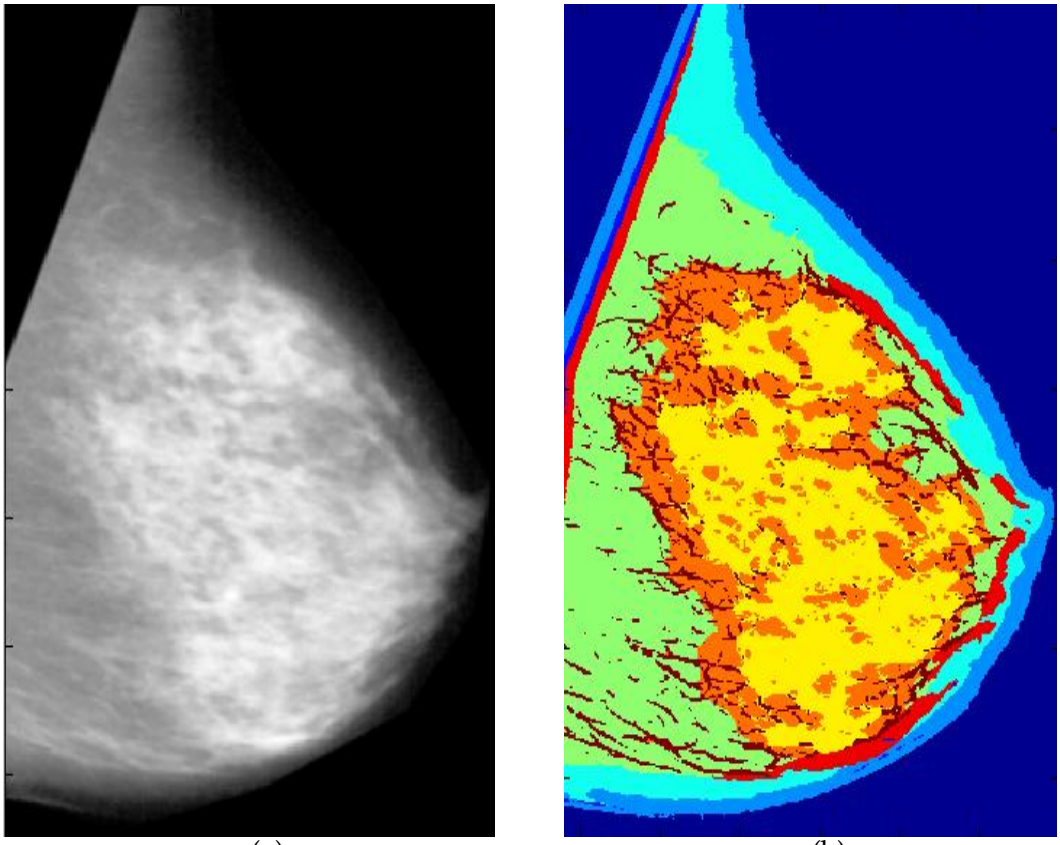
The features and number of clusters were found in a heuristic manner. The lack of ground truth on the tissue segmentation makes it impossible to directly construct a criterion function to use for standard feature selection algorithms.

First, a clustering describing the anatomical composition of the breast tissue, as depicted on the mammogram, is investigated. Subsequently, the condition that the clustering should relate directly to the composition of dense and non-dense tissue is relaxed to allow for other types of clustering.

Using eight clusters and features based on 0th and 2nd order derivatives at different scales and a set of vesselness features [53] lead to a promising combination based on visual inspection. An example of this clustering is shown in Figure 4.1.

Relaxing the condition that the clustering should be visually pleasing lead to considerations on the nature of a suitable feature space for the particular setting of x-ray mammography. Certain properties are desirable, specifically invariance to transformations, which do not relate to the tissue structure, and low noise sensitivity.

Looking at historical, multi-site data, one would like features to be invariant to the monotonic transformations caused by variations in film material, development and digitization. The presence of noise in the images



(a) (b)

Figure 4.1: Clustering using eight clusters and features based on 0th and 2nd order derivatives at different scales and a set of vesselness features. a) Segmented input mammogram; (b) Clustered mammogram

means one cannot rely on pure analysis of isophotes and some robustness of the features with respect to noise is needed.

This inspired an approach using features based only on 2nd order structure, disregarding visual appearance and ensuring the desired invariance properties. Since ninety percent of breast cancers arise from the ductal and lobular glands [3] we chose to investigate features describing the local elongatedness or stripiness.

**Stripiness features based on the Hessian**

The proposed features are invariant to affine intensity transformations of the image and, in addition, point noise robustness is provided through convolution with a Gaussian kernel. For every pixel in the breast tissue, features based on eigenvalues of Hessian at three scales are determined.

The Hessian at scale $s$ is defined by

$$H_s(I) = G_s * \begin{bmatrix} \frac{\partial^2 I}{\partial x^2} & \frac{\partial^2 I}{\partial x \partial y} \\ \frac{\partial^2 I}{\partial y \partial x} & \frac{\partial^2 I}{\partial y^2} \end{bmatrix}$$

where $G_s$ denotes the Gaussian at scale (standard deviation) $s$. This is implemented by analytical derivation of the Gaussian prior to convolution using the fact that $G * \partial I = I * \partial G$ [54]. The numerical implementation takes advantage of the Fast Fourier Transform and the convolution is carried out through the Fourier domain [55]. The features used are given by the quotient:

$$q_s = \frac{|e_1| - |e_2|}{|e_1| + |e_2| + \epsilon}$$

where $e_1$ and $e_2$ are eigenvalues of the Hessian at scale $s$, $e_1 > e_2$, and $\epsilon$ is a small positive number to avoid numerical instability by near-zero division in the planar regions of an image where $e_1 \approx e_2 \approx 0$.

The ratio, $q_s$, is related to the elongatedness of the image structure at the point $(x, y)$ at the scale $s$, hence the "stripiness" reference. The three scales used to determine the Hessian are 1, 2 and 4 mm. This specific choice was made to represent local structure.

**Short comparison of the two approaches on the HRT data**

The visually pleasing clustering did not lead to a significant separation of HRT and Placebo patients based on cluster areas. We suspect the reason is that although some robustness with respect to noise was provided through convolution with a Gaussian kernel the 0th and 2nd order derivatives vary proportionally with intensity changes.

Ideally all the steps in the acquisition pipeline are calibrated and can be corrected for in the end. If this is the case a framework of working with a surface of 'interesting tissue' denoted $h_{int}$ has been proposed by Brady and Highnam [37]. Unfortunately, we have no calibration data from the film acquisition or from the scanning process, and the intensity values vary too much between the different images to be relied upon, directly, for separation.

A highly significant separation between HRT and placebo patients was achieved when employing the stripiness features and using four clusters. As described in the introduction, the overall aim was to apply promising methods to data as soon as possible. This was to get a better understanding of the nature of the mammographic data and the potential of various methods. Therefore further selection and experimenting with clusters and

features were postponed until the validity of the method, using the newly found features, was determined.

### 4.1.1 The mammographic pattern measure

To generate a representative collection of features, 10,000 features are sampled from random positions in every image in the data set. Subsequently, a $k$-means clustering [50] is applied and a nearest mean classifier [50] is constructed from the clustered data. This classifier is applied to all the images providing a new data set consisting of images with labelled pixels. The specific choices of k-means clustering and nearest mean classification were made for ease of interpretation and implementation. Since we are exploring the data it is difficult to say beforehand that a certain classifier or clustering procedure is better when it comes to classification performance. The number of clusters, $k$, was decided after evaluating the separation of the HRT and Placebo groups for different values of $k$. The performance did not vary hugely for $k > 2$, and $k = 4$ was selected as the best candidate. Features and corresponding clustering are illustrated in Figure 4.2.

The mammographic pattern score is determined as a linear combination[1] of the relative areas of the classes in the breast tissue. A linear Fisher discriminant analysis [56] was used find the exact linear combination giving the best separation of the placebo and the HRT groups. From the discriminant analysis it followed that a simplified linear combination of only two area measurements provides a near optimal scoring. The separation using these two areas is illustrated in Figure 4.3. Taking $\beta$ to correspond to the blue area as indicated in Figure 4.2 (d) and $\gamma$ to be the corresponding green relative area, per image, the score is given as $\beta - 2\gamma$. Clustering and classification were implemented using the PRTOOLS [57] library for Matlab.

## 4.2 Evaluation of the unsupervised method

It is natural to compare the new pattern recognition based method (PR) to the threshold based (TH) used earlier. Therefore the unsupervised mammographic pattern score was computed on the data presented in chapter 3.

ROC curves of the PR score and the TH density are compared in Figure 4.4. It shows that the PR score is better overall at classifying the patients into HRT and placebo groups, measured as area under the curve. However, a few early false positives for the PR score makes the TH score best until a false positive rating of about 0.1. In terms of p-values the two measures

[1] Using a quadratic classifier to determine the combination of the relative areas was investigated, but gave comparable results

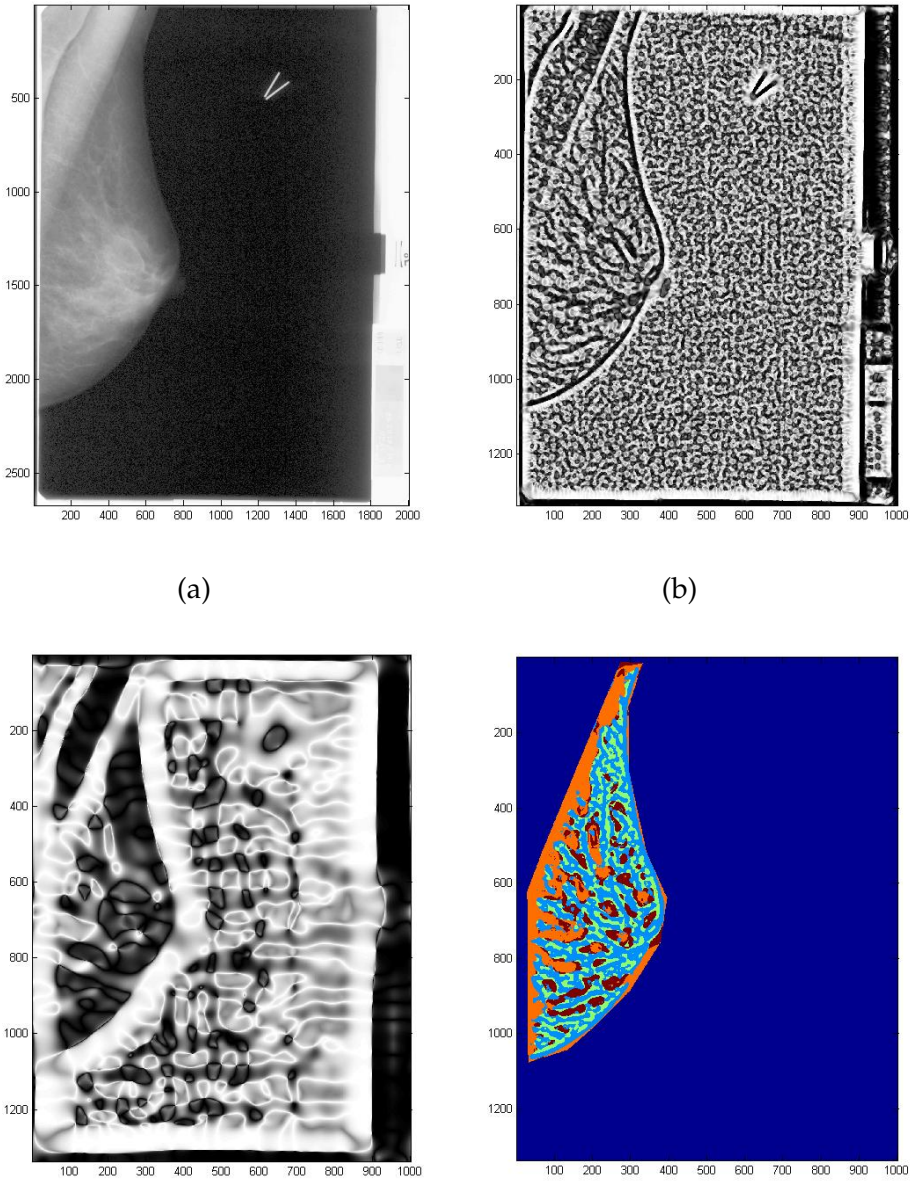



Figure 4.2: Illustration of features and clustering. (a) Input mammogram; (b) and (c) show the smallest and largest scale feature images respectively; d) The tissue clustering used to compute the mammographic pattern score

both perform very well. When checking if the density means of the HRT group in 2001 is significantly higher than in 1999, both scorings have $p <$

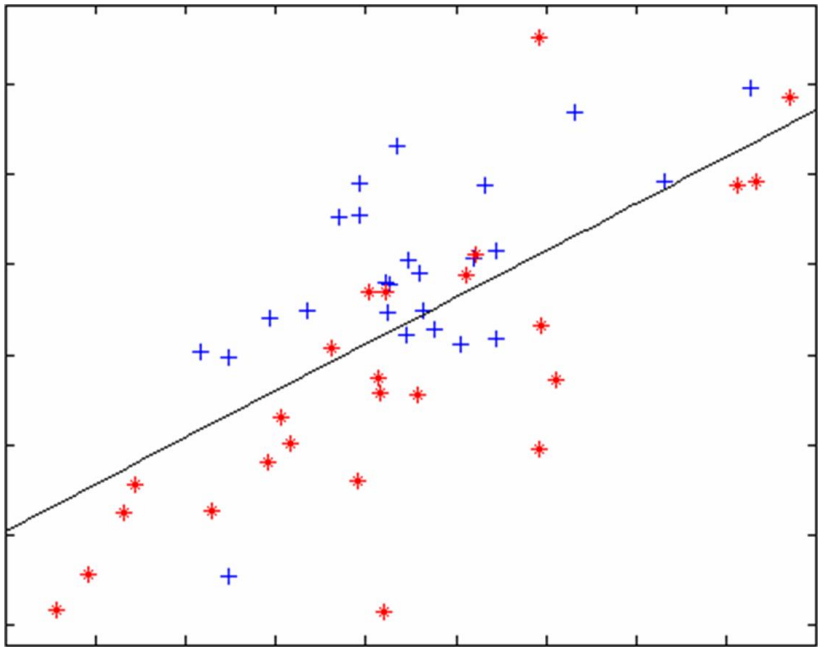


Figure 4.3: Best linear separation of the HRT and Placebo groups using Fisher discriminant. The two axes show the change of $\beta$ (x-axis) and $\gamma$ (y-axis) from 1999 to 2001. + indicate placebo and * HRT.

0.001. Combining the two measures gave no significant improvement.

The evaluation of the density measure is done in a leave one out approach. The linear classifier is finding the best combination of $\beta$ and $\gamma$ for each case based on the N-1 remaining cases and used to predict if the remaining image is from an HRT or a placebo patient. The combinations were all close to $\beta - 2\gamma$, and this rule was fixed after these initial results.

### 4.2.1 Further evaluation

It is investigated how this new measure and the automatic threshold measure compares to two state of the art density assessments, BI-RADS [20] and interactive percentage density [11]. The focus is not on quantifying absolute density changes, but on evaluating the separation between patients in the HRT study. For an accurate and sensitive method a low p-value is expected. In addition to the radiologist readings, the study is expanded by including 15 more patients in each group so there is data from 40 placebo and 40 HRT patients.

In the experiments the reading radiologist was blinded with respect to treatment and the patients were presented in random order. The same radiologist made all readings.

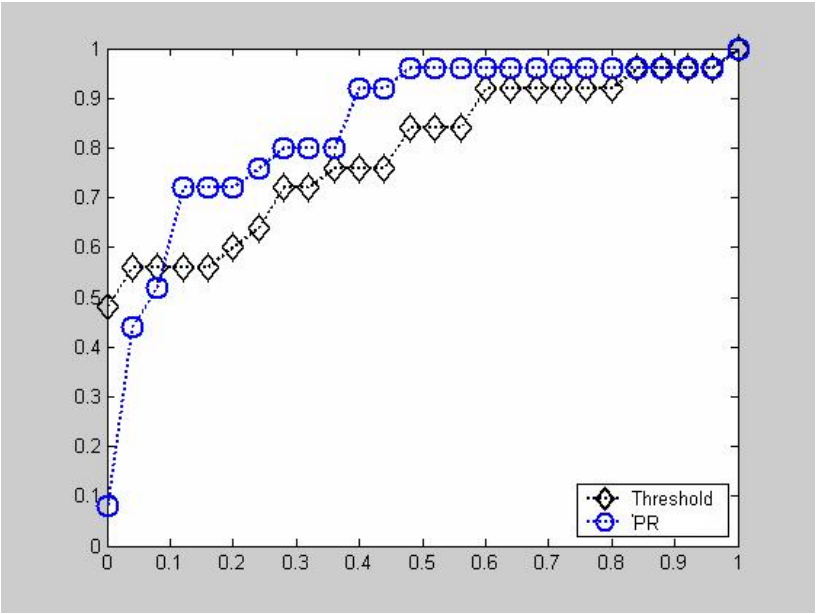



Figure 4.4: ROC curves for the PR score and TH density measures with an are under the curve of 0.82 and 0.76 respectively

### BI-RADS

Breast imaging reporting and data system (BI-RADS) is the four category scheme proposed by the American College of Radiology [20]. The BI-RADS categories are: 1) Entirely fatty; 2) Fatty with scattered fibroglandular tissue; 3) Heterogeneously dense; 4) Extremely dense. A trained radiologist assigns the mammogram to one of these categories based on visual inspection. It is included here since it is widely used both in clinical practice and for automated and computer-aided approaches [22].

### Interactive threshold method

The reading radiologist determines an intensity threshold using a slider in a graphical user interface. She is assisted visually by a display showing the amount of dense tissue corresponding to the current slider position. The system is similar to the approach proposed by Byng et al. [11] and has been used in several clinical studies [22]. The density is defined as the ratio between segmented dense tissue and total area of breast tissue. This specific implementation was made in Matlab and is illustrated in Fig. 4.5.

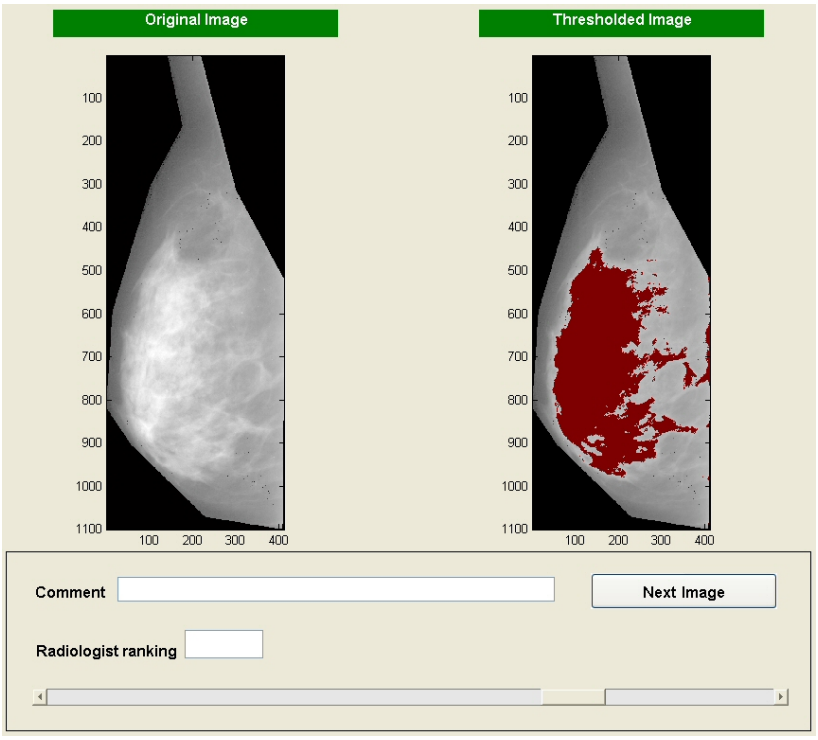



Figure 4.5: Screen dump of the implemented percentage density tool.

### 4.2.2 Results

In the analysis of the density measurements we divide the patient populations into four subgroups. HRT at beginning of study (H99), HRT at end of study (H01), placebo at beginning of study (P99), and placebo at end of study (P01). We do t-tests on four subgroup combinations. Unpaired t-tests on P99 vs. H99 and P01 vs. H01. Paired t-tests on P99 vs. P01 and H99 vs. H01. The zero hypothesis is in each case that the two tested subgroups have identical density means, and the alternative hypothesis that they have different density means.

Table 4.1 displays the p-values for the different tests. No method separates the P99 and P01 groups significantly, and more importantly the P99 and H99 groups, confirming successful randomization of the trial. All methods are able to separate H99 and H01 to a very high degree of significance. Although highly significant the longitudinal separation of the PR score seem a bit lacking compared to the other three. Only the interactive percentage density and the PR score significantly separate H01 from P01.

Table 4.1: p-values for the different methods and tests.

| Method \ Test | P99 vs P01 | H99 vs H01 | P99 vs H99 | P01 vs H01 |
|---|---|---|---|---|
| BI-RADS | 0.3 | $< 0.001$ | 0.3 | 0.1 |
| Interactive TH | 1 | $< 0.001$ | 0.8 | 0.02 |
| Automatic TH | 0.07 | $< 0.001$ | 0.8 | 0.2 |
| PR score | 0.9 | 0.004 | 0.9 | 0.02 |

## 4.3 Discussion and conclusions

The interactive threshold shows better capability to separate the HRT patients from the placebo patients at end of study, than the categorical BI-RADS methodology. This was expected due to the continuous nature of the threshold measure. Also, the computer-aided measure has been reported to have a lower intra-observer variability than the BI-RADS measure [21]. For the automated methods, the PR score displays a similar increase in performance when compared to the automatic threshold. We think, that this is mainly due to the intensity invariance of the stripniness measure.

Contemporary to this development of an invariant density measure is the work by Pan et al. [58] who use monogenic signal processing to achieve a robust set of features. While the features are presented in the context of breast boundary segmentation, the methodology seems promising for other, related tasks such as density estimation.

It is a drawback of our proposed mammographic pattern score that it is harder to interpret visually than the threshold density. The change in brightness detected by the TH density is easy to understand compared to the change in clustering and corresponding change in PR score, which is subtle and difficult to interpret directly. Based on a preliminary investigation [15], using artificial images composed of sinusoid combinations, it appears that becoming more isotropic leads to an increased HRT likelihood. It also seems that tending towards a wave length of about 15 pixels (corresponds to 3 mm) from a smaller wave length also increases the PR score. Since the correspondence between real mammograms and sinusoid images is not straightforward these results merely give an indication of the kind of changes the classifier picks up. A more precise quantitative description of the changes, and a discussion of these with physicians, are needed to get a qualitative understanding of the structural changes detected in the HRT group.

In conclusion, we have shown that unsupervised clustering of mammograms based on the quotient of Hessian eigenvalues at three scales can

be used to differentiate between patients receiving HRT and patients receiving placebo. The proposed mammographic pattern score is an automated method, able to quantify the effect of HRT as structural changes in the breast tissue. To our knowledge the Hessian eigenvalues have not been used in connection with density in any previous work.

So far, the only step using the knowledge of patient labels is the determination of the best area combination using the linear discriminant. The information about patient labels may be applied earlier in the process, and a supervised framework, applying this knowledge using the same stripiness features, is presented in the next chapter.

# Chapter 5

# Supervised method

In this chapter we present a supervised approach and demonstrates that it can be trained to detect changes due to aging and HRT. Because of the supervised machine learning approach employed, the method can be easily adapted to the detection of other mammographic changes.

## 5.1 Introduction

The aim of the presented work is to provide a framework for obtaining more accurate and sensitive measurements of breast density changes related to specific effects. Given effect-grouped patient data, we propose a statistical learning scheme providing such a non-subjective and reproducible measure and compare it to the BI-RADS measure and a computer-aided percentage density.

Several approaches to other automatic methods for assessing mammographic breast density have been suggested [42, 38, 32, 59, 45]. All of these aim at reproducing the radiologist's categorical rating system or at segmenting the dense tissue to get a percentage density score. Our approach differs from existing methods in mainly three ways

1. Breast density is considered a structural property of the mammogram, that can change in various ways explaining different effects.

2. The measure is derived from observing a specific effect in a controlled study.

3. The measure is invariant to affine intensity changes.

It is noted that we do not aim at measuring what is traditionally called breast density, i.e., the relative amount of fibroglandular tissue. Since the term mammographic density is most often used for this type of measure, we have decided to use "mammographic pattern" to describe more general

properties of the mammogram. We mean to convince the reader of the fact that mammographic changes can perceived as a structural matter that may be accessed ignoring the actual brightness of the images and that it changes differently under the physiological processes of aging and HRT.

The following section, Section 5.2, introduces the medical study that produced the images used in this investigation. Subsequently, Section 5.3 describes the two standard methods and the new supervised method in detail. Section 5.4 contains a description of the experimental setup and results. Section 5.5 consists of discussions and conclusion.

## 5.2 Materials

The data used in this work is from a 2-year randomized, double-blind, placebo-controlled clinical trial, in which the participants received either 1 mg 17$\beta$-estradiol continuously combined with 0.125 mg trimegestone (n=40), or placebo (n=40) for 2 years. At entry into the study, women were between 52 and 65 years of age[1], at least 1 year postmenopausal with a body mass index less than or equal to 32 kg/m$^2$.

Breast images were acquired at the beginning ($t_0$) and the end of the 2-year treatment period ($t_2$) using a Planmed Sophie mammography X-ray unit. The images were then scanned using a Vidar scanner to a resolution of approximately 200 microns with 12 bit gray-scales. Delineation of the breast boundary on the digitized image was done manually by an expert using 10 points along the boundary connected with straight lines. Only the right mediolateral oblique view was used, since it has been shown previously that a reliable measure of the breast density can be assessed from any one view [47]. We denote the patient groups P0, P2, H0, and H2 for placebo and treatment at $t_0$ and $t_2$ respectively.

## 5.3 Methods

For both methods involving human interaction, the reading radiologist was blinded with respect to treatment and the images were presented in random order. The same radiologist made all readings.

### 5.3.1 BI-RADS

Breast imaging reporting and data system (BI-RADS) is the four category scheme proposed by the American College of Radiology [20]. The BI-RADS

[1]Placebo and HRT groups are age-matched in the sense that their mean ages are not significantly different

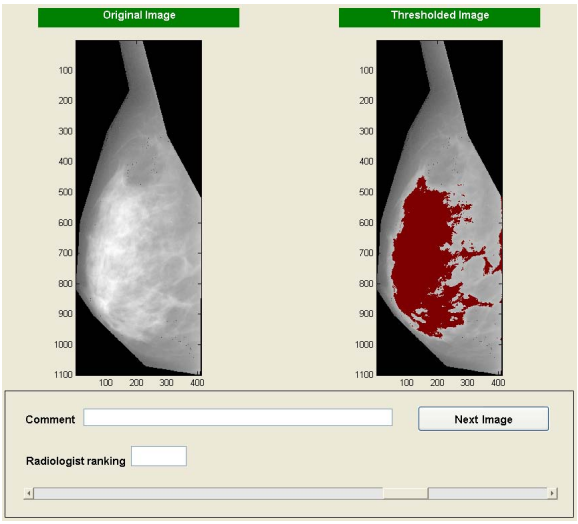



Figure 5.1: Screen dump of the implemented percentage density tool.

categories are: 1) Entirely fatty; 2) Fatty with scattered fibroglandular tissue; 3) Heterogeneously dense; 4) Extremely dense. A trained radiologist assigns the mammogram to one of these categories based on visual inspection. It is included here since it is widely used both in clinical practice and for automated and computer aided approaches [22].

### 5.3.2 Interactive threshold method

The reading radiologist determines an intensity threshold using a slider in a graphical user interface. She is assisted visually by a display showing the amount of dense tissue corresponding to the current slider position. The system is similar to the approach proposed by Yaffe [11] and has been used in several clinical trials [22]. The density is defined as the ratio between segmented dense tissue and total area of breast tissue. Our implementation was made in Matlab and is illustrated in Fig. 5.1.

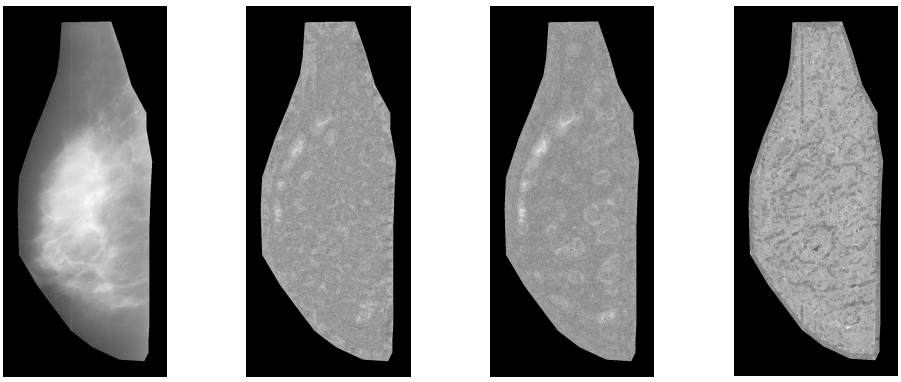
### 5.3.3 The supervised approach

Our mammographic pattern measure is derived by training a pixel classifier on subsets of images from the available data. These subsets are chosen to represent the potential differences in patterns to be detected by the method. As an example, one subgroup may be the H2 images from hormone treated patients and the other the P2 images from the placebo group.

Most often, as in our case, the pixel classification would be based on local features that describe the image structure in the vicinity of every pixel to be classified. Generally, the features extracted per pixel will exhibit large

similarity for every image even though they may come from two different subgroups of images. Therefore, for individual pixels, it will be difficult to decide to which of the subsets it belongs. Fusing all weak local decisions, however, into a global overall score per image ensures that sufficient evidence in favor of one of the two groups is accumulated and allows for a more accurate decision.

In this work, a simple fusion strategy is employed. After every pixel has been provided with a posterior probability by the classifier, the average probability per pixel in the image is determined. This mean is then taken as the final score. Obviously, several other fusion schemes are possible (see e.g. [60]), but we do not necessarily expect benefit from these. An example of a mammogram with corresponding pixel probability maps is shown in Fig. 5.2. Below follows a more precise description of the features and a description of the various subgroups used to train the classifiers.



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

Figure 5.2: Mammogram from the data set (a); pixel classification result using the classifiers HRTC, HRTL and AGE respectively (b), (c), (d)

**Features.**

A specific three dimensional feature space is used since a previous study found these features to be associated with breast density [13]. These features are invariant to affine intensity transformations of the image[2] and, in addition, point noise robustness is provided through convolution with a Gaussian kernel. For every pixel in the breast tissue, features based on eigenvalues of Hessian at three scales are determined. The Hessian at scale

[2]Almost, that is. Due to the addition of the small constant $\epsilon$ in the denominator, strict intensity invariance is not attained.

$s$ is defined by

$$H_s(I) = G_s * \begin{bmatrix} \frac{\partial^2 I}{\partial x^2} & \frac{\partial^2 I}{\partial x \partial y} \\ \frac{\partial^2 I}{\partial y \partial x} & \frac{\partial^2 I}{\partial y^2} \end{bmatrix}$$

where $G_s$ denotes the Gaussian at scale (standard deviation) $s$. This is implemented by analytical derivation of the Gaussian prior to convolution using the fact that $G * \partial I = I * \partial G$ [54]. The scales used are 1, 2 and 4 mm. The features used are given by the quotient:

$$q_s = \frac{|e_1| - |e_2|}{|e_1| + |e_2| + \epsilon}$$

where $e_1$ and $e_2$ are eigenvalues of the Hessian at specific scale $s$ and $e_1 > e_2$, and $\epsilon$ is a small positive number ($\epsilon = 10^{-5}$) to avoid numerical stability problems in the planar regions of an image where $e_1 \approx e_2 \approx 0$. This ratio is related to the elongatedness of the image structure at the point $(x, y)$ at the scale $s$.

**Subgroups and classifiers.**

Three combinations of subgroups are used for classifier training and tested in the experiments conducted subsequently:

`HRTL`
: Subsets H0 and H2 are used to capture the effect of HRT. There is also an effect of aging, but it is expected to be much lower than that of HRT. The trained classifier is referred to as `HRTL` (longitudinal).

`HRTC`
: Subsets P2 and H2 are used to capture the effect of HRT. Separation between classes is expected to be lower, since inter-patient biological variability is diluting the results. The trained classifier is referred to as `HRTC` (cross-sectional).

`AGE`
: The baseline population (P0 and H0) is stratified into three age groups, and the first and last tertile are used to capture the effect of age. The second tertile is used as control population. The trained classifier is referred to as `AGE`.

In each case every pixels receives a label based on the subgroup it belongs to and a $k$ Nearest Neighbours ($k$NN) classifier ($k = 100$) is trained using this data to classify pixels from the two classes. For every pixel a posterior probability can be determined of belonging to one of the two classes.

This posterior is simply determined as the number of $k$ neighbors that are assigned to the one class divided by $k$, which is a standard procedure [50]. The use of this powerful, non-parametric classifier is justified by the large number of pixels and low dimensionality of the feature space.

We implement our classifier using an approximate $k$NN framework developed by Arya and colleagues [61]. The approximate classifier is in principle a $k$NN-classifier, but allows for faster computations if an error is tolerated. The search algorithm returns $k$ points such that the ratio of the distance between the $i$th approximated nearest point ($1 \leq i \leq k$) and the true $i$th nearest neighbor is at most $1 + \epsilon$. We found that any $\epsilon \leq 2$ could be used without degrading the results noticably.

Although there are a large number of pixels at our disposal, the number of patients is rather limited. For this reason, the data is not split up into a single training and a test set. Instead the classifier is trained on all but a pair of images (one image from each class) and pixel probabilities are computed for this pair using the trained classifier. This is repeated until all pixel probabilities for all images are computed. What is basically conducted is a leave-one-out procedure [50] on the image and not the pixel level, with the slight additional modification to leave-two-out since leaving one sample from class $A$ out introduces a bias for belonging to class $B$. This is especially needed in our setting where the number of samples, i.e., patients, is relatively low (80 for the HRT classifiers and 56 for the age classifier). Feature vectors are extracted from 10,000 randomly selected pixels within the breast region in each image.

## 5.4 Experimental setup and results

In the experiments, $k$ is set to 100 as initial pilot experiments indicated that this gives good results. No additional tuning was performed. In order to judge the impact of varying $k$, we checked its influence on the `HRTC` classification. The results are given in Figure 5.3, indicating that our initial choice for $k$ is indeed fine.

The experiments serve to answer two questions. How does the separation of the hormone treated subpopulation, H2, compare to the same patients at baseline, H0, and the control population who received placebo, P2, for the different measures? And, can any of the measures detect the aging of the placebo group by separating P2 and P0? Statistical t-tests are used to test for significance in the separation and resulting $p$-values make a comparison of methods possible.

Table 5.1 shows $p$-values for all combinations of methods and relevant pairs of groups. The first two columns are paired two-sided t-tests, while the last two columns are unpaired. In Fig. 5.4 the pattern changes are shown using the three different training strategies together with the BI-

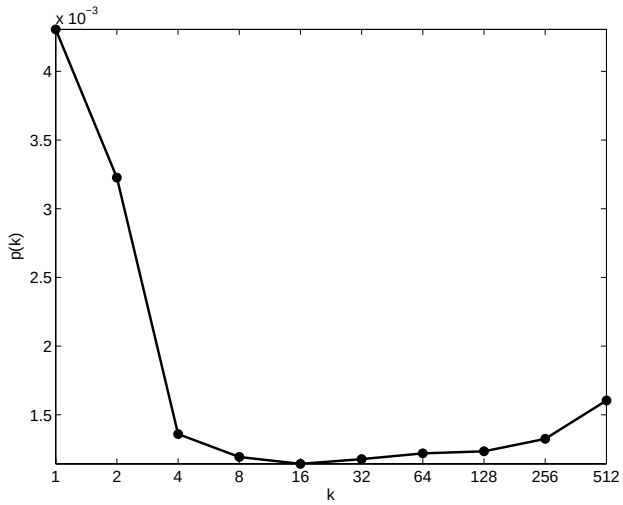



Figure 5.3: $p$-values for H2 versus P2 separation as function of $k$ using the HRTC classifier.

RADS scores and the percentage density. The figure allows for a qualitative comparison of the methods by showing the progressions of the HRT and placebo groups combined with standard deviations of means.

Table 5.1: $p$-values for the different methods and tests. Thresholding is abbreviated TH.

| Method\Test | P0 vs. P2 | H0 vs. H2 | P0 vs. H0 | P2 vs. H2 |
|---|---|---|---|---|
| BI-RADS | 0.3 | $< 0.001$ | 0.3 | 0.1 |
| Interactive TH | 1 | $< 0.001$ | 0.8 | 0.02 |
| HRTL | 0.08 | $< 0.001$ | 0.7 | 0.003 |
| HRTC | 0.4 | 0.003 | 0.7 | 0.001 |
| AGE | 0.004 | 0.4 | 0.8 | 0.07 |

Table 5.2 shows ROC areas for the task of separating P2 vs. H2 and the differences in ROC areas and bootstrapped statistics. Although the standard deviations are quite large, the analysis indicates that, HRTC outperforms all other approaches. HRTL is second best, though closely followed by interactive thresholding. All methods perform better than BI-RADS. The ROC curves of the four measures are shown in Fig. 5.5.

Fig. 5.6 shows that the differences between P0 and P2 indicated by the

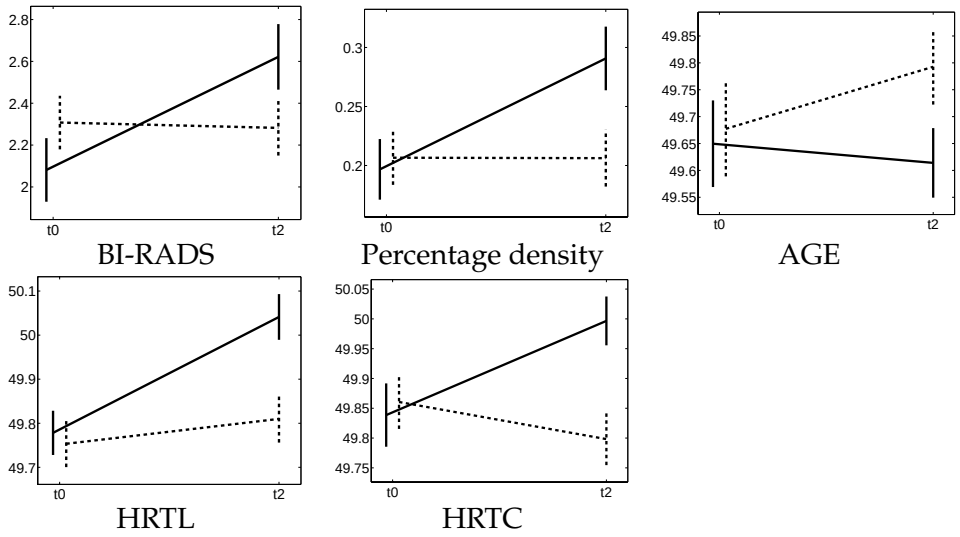



Figure 5.4: Longitudinal progression of the different measures. The placebo group is indicated with a dashed line; HRT group by a solid. Vertical bars indicate the standard deviation of the mean of the subgroups at $t_0$ and at $t_2$.

Table 5.2: Top: ROC areas and standard deviations. Bottom: Differences in ROC areas and their standard deviations

| Measure\ROC stats | AUC | std |
|---|---|---|
| BI-RADS | 0.61 | 0.06 |
| Interactive TH | 0.66 | 0.063 |
| HRTL | 0.69 | 0.06 |
| HRTC | 0.71 | 0.06 |
| TH - BI-RADS | 0.056 | 0.028 |
| HRTL-TH | 0.028 | 0.067 |
| HRTC-TH | 0.049 | 0.055 |
| HRTC-HRTL | 0.021 | 0.056 |

AGE classifier is indeed an age related effect and not a general difference in image appearance at $t_0$ and $t_2$. The entire baseline population is again stratified into three age groups and the AGE measures show an increasing trend with increasing age. The means values of the first and last tertile are significantly different ($p = 0.015$).

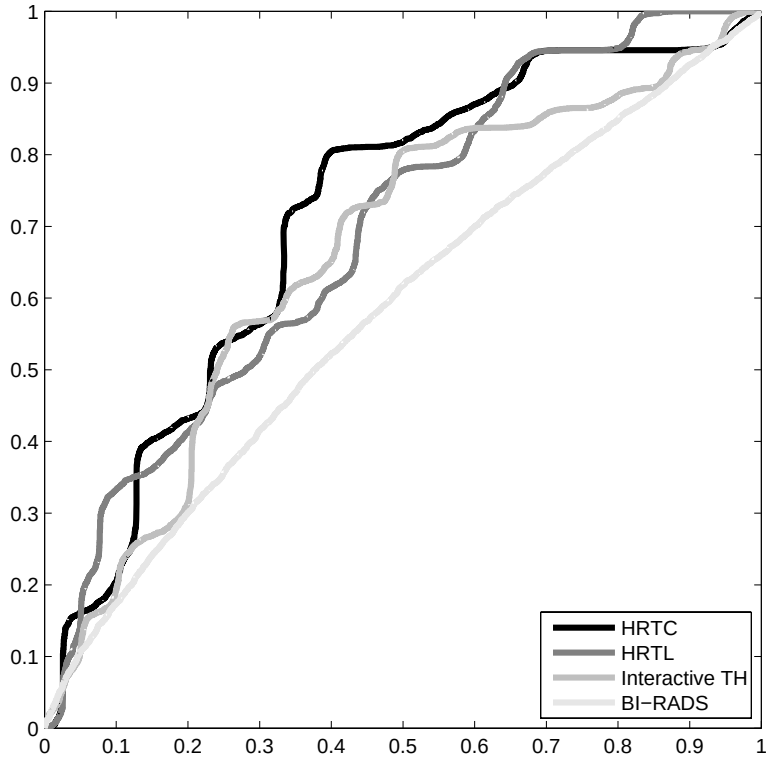



Figure 5.5: ROC curves for the four compared measures separating P2 and H2.

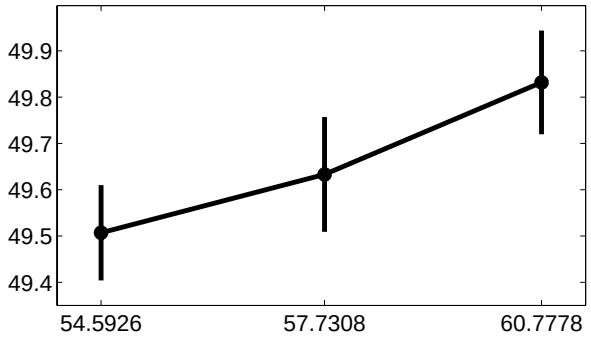



Figure 5.6: `AGE` mammographic pattern as a function of the means of the three age tertiles in the baseline population. Vertical bars indicate the standard deviation of the mean of the corresponding tertile.

## 5.5 Discussion and conclusions

The first observation that should be made is that none of the methods separate the two baseline groups P0 and H0, confirming successful random-

ization. The second immediate observation is that the interactive threshold shows better capability to separate P2 and H2 than the categorical BI-RADS methodology. This might be explained by the continuous nature of the threshold measure making it more sensitive.

For the automatic measures, `HRTL` and `HRTC` performs even better than the percentage density, and `AGE` detects the aging effect in a very significant way as opposed to the currently available techniques, which are unable to detect any meaningful changes.

The inverse appearance of `AGE` and `HRTC` changes on Fig. 5.4 suggests that the age-related mammographic pattern and the HRT-trained pattern change along directions in our Hessian-based feature space that are not orthogonal, but rather somewhat pointing in opposite directions. This behavior is in agreement with density increasing with HRT and decreasing with age [62, 3].

In conclusion, the proposed methodology shows substantial merit as it performs considerably better than both the BI-RADS and the percentage density method, the current state of the art. As shown in this work, the approach can be trained to detect changes due to aging and HRT. These changes might not be interesting in themselves, but because of the supervised machine learning approach employed, the method can be easily adapted to the detection of other mammographic changes.

# Chapter 6

# Supervised framework extended using SFS feature selection

In this chapter we present a framework for incorporating feature selection in our supervised methodology. This framework is applied to a set of data from the Dutch national breast cancer screening program. The presented results demonstrates the ability and potential of including feature selection to improve and specialize measures.

In the two previous chapters, we showed that the stripiness features performed well on HRT data, both in an unsupervised and a supervised setting. Obviously, these features are not expected to perform well in all situations and, generally, the performance of our method may improve by allowing more features. However, indiscriminately adding features will eventually deteriorate the results. One way to cope with this situation is by means of a feature selection strategy.

A somewhat related study was carried out by Huo et al.[63], where 14 image features are related to measures of breast cancer risk. They employ linear discriminant analysis is to identify features that are useful in differentiating between low-risk women and BRCA1/BRCA2-mutation carriers. Linear regression analysis is employed to identify useful features in predicting the risk, as estimated from the Gail [28] and Claus [64] models. They find that women at high risk tend to have dense breasts and their mammographic patterns tend to be coarse and low in contrast.

The study presented is this chapter differs from the work by Huo et al. in various ways. The main differences are that we investigate local features not global and that we evaluate on a large set of mammograms from women who were actually later diagnosed with cancer versus a similar set of controls. We find local mammographic features, mainly describing the structure around the vertical axis and the position in the breast, which are

indicative of women developing cancer (AUC = 0.70). The feature with the highest association of risk found by Huo et al., histogram skewness, was less indicative (AUC = 0.60).

The following section, Section 6.1, describes the proposed methodology in detail. Subsequently, Section 6.2 introduces the data investigated in this study. Section 6.3 contains a description of the experimental setup and results. Section 6.4 consists of discussions and conclusion.

## 6.1 Methodology

Why not just use all the features we can think of? Well, there is the well-known problem of overfitting to consider. If you use enough features and a powerful classifier it is possible to separate almost anything, but the resulting classifier loses the ability to generalize to new data, since the demand of samples grows exponentially with the dimensionality of the feature space [60]. This problem is also known as the "curse of dimensionality" and implies that only a limited number of features may be used effectively, depending of the number of samples in your data set.

The goal of feature selection in pattern recognition is to select the most discriminative features from a given feature set to improve classification performance. Through the process of feature selection, we can potentially accomplish the following:

- Improved classification performance.
- Better understanding of the relationship between features and classes.
- Less computing resources needed for building (and, depending of type, running) the classifier.

The first two improvements are of special interest to us, since we are ultimately interested in identifying the features most indicative of breast cancer risk.

The aim of feature selection can be stated more formally as follows. Given a feature set $\mathbf{F}$, we construct a classifier with a recognition rate $R(\mathbf{F}')$ as a function of the selected features, $\mathbf{F}'$. The goal of feature selection is to select the subset $\mathbf{F}'$ of $\mathbf{F}$ such that $R(\mathbf{F}') > R(\mathbf{T})$, where $\mathbf{T}$ denotes all possible subsets of $\mathbf{F}$. Several choices are available for quantifying the recognition rate, including specificity, sensitivity, area or volume overlap of a segmentation task, and area under ROC curve to name a few. Which choice to make depends on the application. It should be noted that, independent of choice, it is important to evaluate the recognition rate on data that are independent of the training data. This is typically done, either by splitting the data up in train and test sets or use a leave-one-out approach for evaluating the recognition rate[60].

Among other things, due to the combinatorial explosion, there is generally no efficient way to determine the optimal feature set and we have to resort to suboptimal approaches, which typically determine a suboptimal feature set. For an introduction to and overview of the different ways of approaching this problem the reader is referred to [65]. In our current approach, we employed a basic sequential forward selection method, as originally proposed by Whitney [66]. It is one of the commonly used heuristic methods for feature selection and involves the following steps:

1. Select the first feature that has the highest recognition rate among all features.
2. Select the feature, among all unselected features, that gives the highest recognition rate together with the selected features.
3. Repeat the previous step until you have reached a preset number of features, until the recognition rate exceeds a preset threshold, or until all features are selected.

### 6.1.1 Features

In addition to the stripiness features previously presented, we propose a set of position features based on a distance map of the breast boundary. Two additional types of features are considered for providing a large set of descriptive features selectable in the feature selection process. One is the set of invariant, differential features proposed by Romeny et al.[1] that, in principle, describe all local intrinsic properties of a scalar image at a fixed level of resolution. The other is the set of local, partial derivatives up to order $n$, commonly referred to as the $n$-jet. The jets are useful descriptors of local image structure, shown to be related to the processing of the visual system [67].

#### Polynomial invariants

The gauge coordinate frame $(v, w)$ is defined such that $w$ is everywhere along the gradient direction and $v$ tangential to the isophote. These two directions are always perpendicular to each other and form a local coordinate frame. All polynomial expressions in $(v, w)$ are invariant under orthogonal transformations [1]. As one feature set we test all non-singular polynomial invariants up to third order resulting in 8 features per scale (Table 6.1).

#### 3-jet features

The other tested feature set is the 3-jet consisting of all partial derivatives up to third order. This gives 10 features per scale. In calculating both polynomial invariants and 3-jet features, we define the partial derivative of the

Table 6.1: List of non-singular polynomial invariants up to third order expressed in gauge coordinates [1].

| Order | 1 | 2 | | | 3 | | | |
|---|---|---|---|---|---|---|---|---|
| Gauge | $I_w I_w$ | $I_{vv} I_w^2$ | $I_{vw} I_w^2$ | $I_{ww} I_w^2$ | $I_{vvv} I_w^3$ | $I_{vvw} I_w^3$ | $I_{vww} I_w^3$ | $I_{www} I_w^3$ |

image, $I$, at scale, $s$, as

$$I_{xs} = G_s * \frac{\partial I}{\partial x}$$

where $G_s$ denotes the Gaussian with standard deviation $s$. This is implemented by analytical derivation of the Gaussian prior to convolution using the fact that $G * \partial I = I * \partial G$ [54]. The numerical implementation takes advantage of the Fast Fourier Transform and the convolution is carried out through the Fourier domain[55].

Both large feature sets are based on differential features related to image structure and the main difference is the rotational invariance provided by the invariant features. Only the best performing of the two sets are analysed in detail together with the stripness features. We use scales 1, 2, and 4 mm based on previous findings with the stripiness features. In addition, a larger scale of 8 mm is introduced to allow for some larger scale information. This means we are testing 40 jet-features and 32 invariant features.

**Position features**

So far, no information about where in the image a given feature vector was sampled has been available to the classifier. If the changes we are investigating mainly occur in specific regions this knowledge will help reduce noise from changes in unimportant regions. If there are important changes in one region simultaneous with important, but manifested inversely in the conventional features, in another region, this knowledge might improve classification dramatically. Therefore a crude breast coordinate frame is introduced for the feature selection experiments. Three position features are used: 1) Distance to nearest breast boundary implemented as a distance map, 2) Horizontal displacement from center of distance map, and 3) Vertical displacement from center of distance map. A mammogram and corresponding distance map are shown in Figure 6.1. The position features represent a separate category of features and are included in all the experiments.

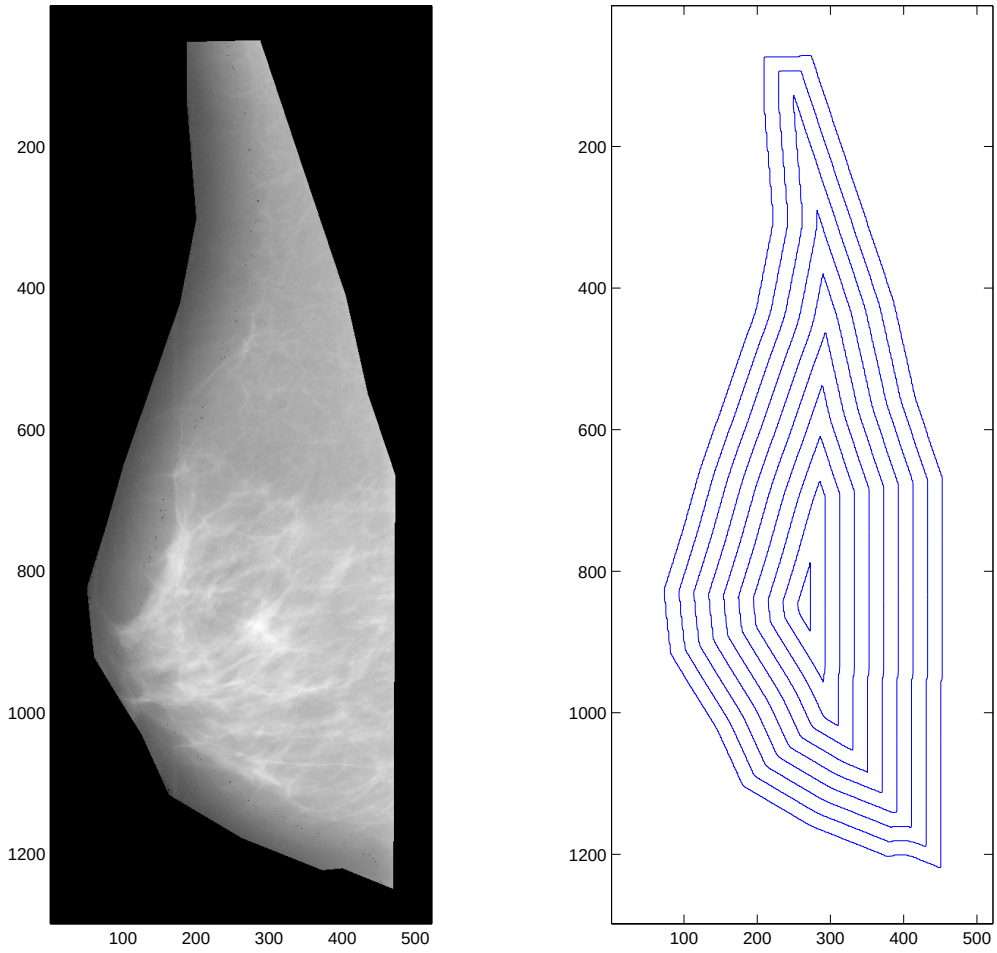
## 6.2 Materials

The investigated mammograms are from the Dutch national breast cancer screening program. The data was originally used to investigate the effect of recall rate on earlier screen detection of breast cancers [68]. Mammograms were collected from a total of 495 women participating in the biennial Dutch screening program. Of these, 250 were chosen as control subjects, and 245 were from women who were diagnosed with breast cancer. The data include screening mammograms from the time of diagnosis and screen-negative mammograms from at least two preceding screening examinations for both cases and controls.

The data set used in this study was formed by selecting the earliest available screen-negative mammograms for all participants. The result is a high risk (100%) group of cases who were diagnosed with breast cancer within 2-4 years, but radiological reading provided no evidence of cancer at this earliest examination and 2 years after, and a low-risk group who were not diagnosed with breast cancer for a minimum of 4 following years. The segmentation of breast tissue was done automatically using techniques presented by Brady and Highnam [37] (breast boundary) and Karssemeijer [38] (pectoral muscle). Subsequently the masks were postprocessed using a morphological opening with af circular structure element with a diameter of 10 mm and the largest component selected as final breast tissue mask to improve the segmentation quality. Only the right mlo views are analysed in these experiments.

## 6.3 Experimental Setup and Results

In evaluating the performance of the classification of a certain feature set, the data is split up in a training and a test set, each consisting of 100 cancer and 100 control patients. Each component of each feature vector is normalized to unit variance across the entire training set. Standard sequential forward selection is used as feature selection algorithm with recognition rate quantified as area under ROC curve (AUC). The classification step is similar to what is described in the previous chapter, apart from the number of features used to represent each image. Machine memory only allowed 1000 feature vectors used per image due to the increase in feature space dimensionality and sample size. An equivalent $k$ is used, modified to reflect the smaller total number of feature vectors in the training set (four times more in the HRT experiments implying that $k = 25$ here).




(a) (b)

Figure 6.1: A mammogram (a) and contour plot of corresponding distance map (b)

### 6.3.1 Invariant features versus $n$-jet

The performance of the two types of general features is investigated in two separate feature selection runs and the best performing type is selected. Using randomly selected train and test sets each consisting of 100 cases and 100 controls, Figure 6.2 shows the performance of SFS with no stopping criterion applied once using the invariance and position features and once using 3-jet and position. The same patients were used for train and test sets

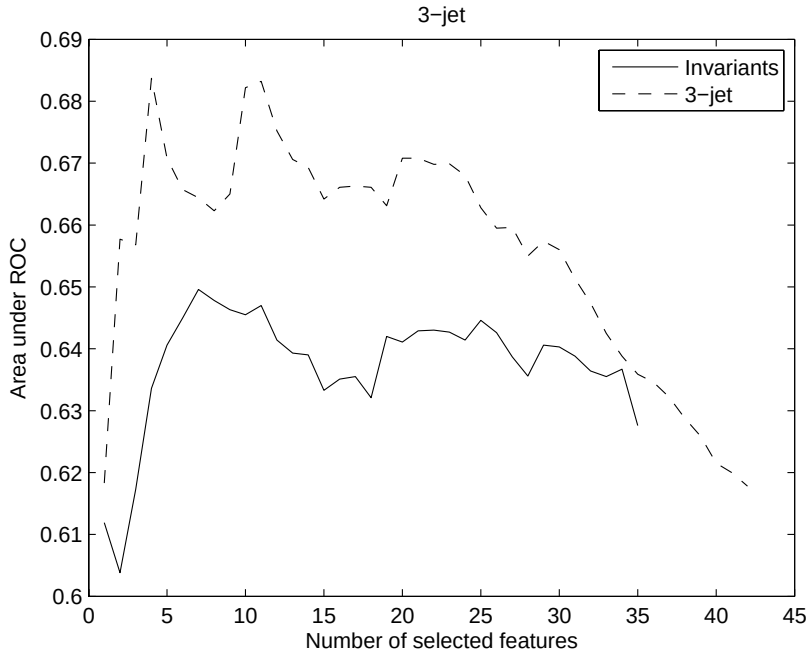



Figure 6.2: ROC area as function of number of selected features. The features where selected using SFS with no stopping criterion.

were used in both cases.

Based on the results (AUC for 3-jet being everywhere larger than for invariants) the 3-jet features are selected and investigated in a large experiment in combination with the stripiness features. Based on Figure 6.2, 10 features are selected as stopping criterion. One might argue that 15 or 20 would be a safer selection number (also including the second local maximum), however we would rather be able to make a clearer inference on the type of features related to risk than potentially getting a, probably, small boost in recognition rate. The first top of the 3-jet ROC in Figure 6.2 ($AUC = 0.6837$) is at four selected features 0.6837 and the second top ($AUC = 0.6832$) at 11 selected features.

### 6.3.2 Gathering feature selection statistics

To gather information on which features are selected 100 SFS runs are calculated for three different setups. Each run uses a new random train and test set. These sets are again each made of 100 cases and 100 controls. First we investigate only 3-jet and position features. Then stripiness features are included as selectable by the SFS algorithm. Finally, it is tested whether forcing SFS to select the three stripiness features improve results.

Table 6.2: The 3-jet features are ordered as follows. This information is needed to read the feature indices of Figures 1.3-5

| Nr. | $\partial x$ | $\partial y$ |
|---|---|---|
| 1 | 0 | 0 |
| 2 | 0 | 1 |
| 3 | 1 | 0 |
| 4 | 0 | 2 |
| 5 | 1 | 1 |
| 6 | 2 | 0 |
| 7 | 0 | 3 |
| 8 | 1 | 2 |
| 9 | 2 | 1 |
| 10 | 3 | 0 |

Images are represented by features from the same 1000 pixels in all experiments and the same 100 randomized train-test sets are used in the three setups making it possible to compare both overall performance and individual runs. Figures 6.3, 6.4, and 6.5 show the results of the 100 runs of $n$-jet, $n$-jet + stripy selectable, and $n$-jet + stripy forced. The features from 1-10 are the 3-jet at scale 1mm, from 11-20 the 3-jet at scale 2mm, from 21-30 at 4mm and 31-40 at 8mm. In order to read the feature numbers the ordering of the 3-jet features is displayed in Table 6.2. Origo of the image coordinate system is in the upper left corner which means that the $x$-direction is vertical and the $y$-direction horizontal. Features 41-43 are the distance to skin line, horizontal displacement, and vertical displacement respectively. Features 44-46 are the stripiness features at scales 1, 2, and 4 mm.

## 6.4 Discussion and conclusion

We have demonstrated the ability and potential of including learning of features to improve and specialize measures. The histograms of selected features in Figures 6.3, 6.4, and 6.5 give some information about the relationship between features and classes, which was one of the potential benefits of feature selection. Though a bit too flat to give a clear picture, it seems that the derivatives in the horizontal direction (2, 4, 7, 12, 14, ... , 37), illustrated in Figure 6.6 and the horizontal and vertical position (42 and 43) are the features most indicative of risk. 0th order features and pure vertical derivatives are very seldom selected. This may be the reason why the 3-jet performed better than the polynomial invariants - the orientation of structure matters.

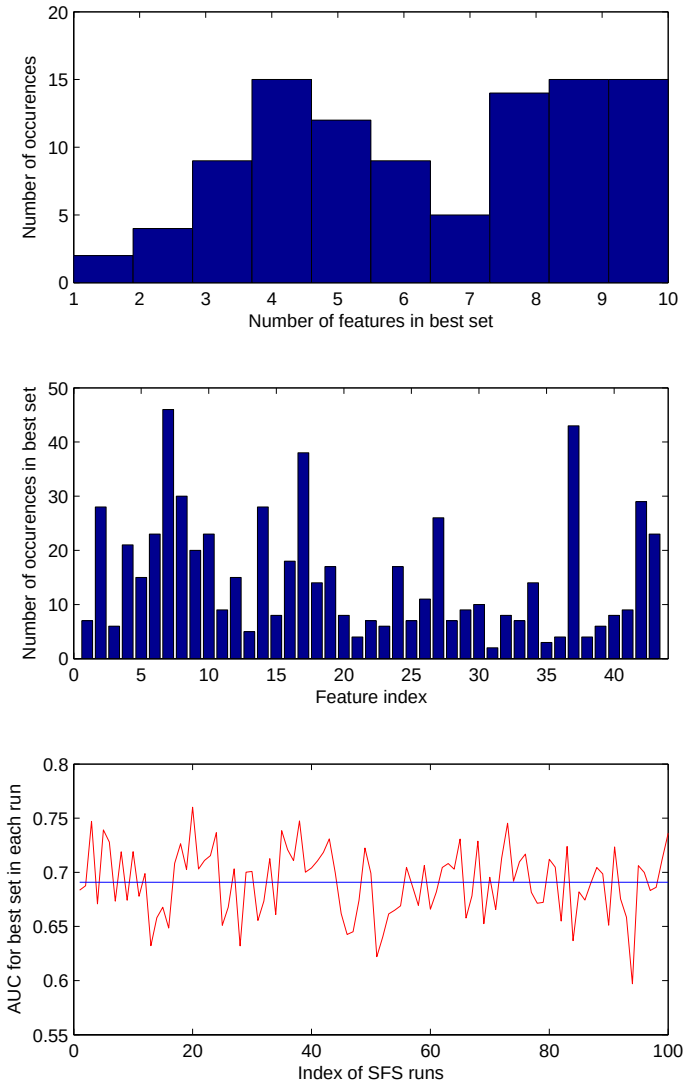



Figure 6.3: Feature selection statistics for only jet features. Average $AUC = 0.69 \pm 0.03$

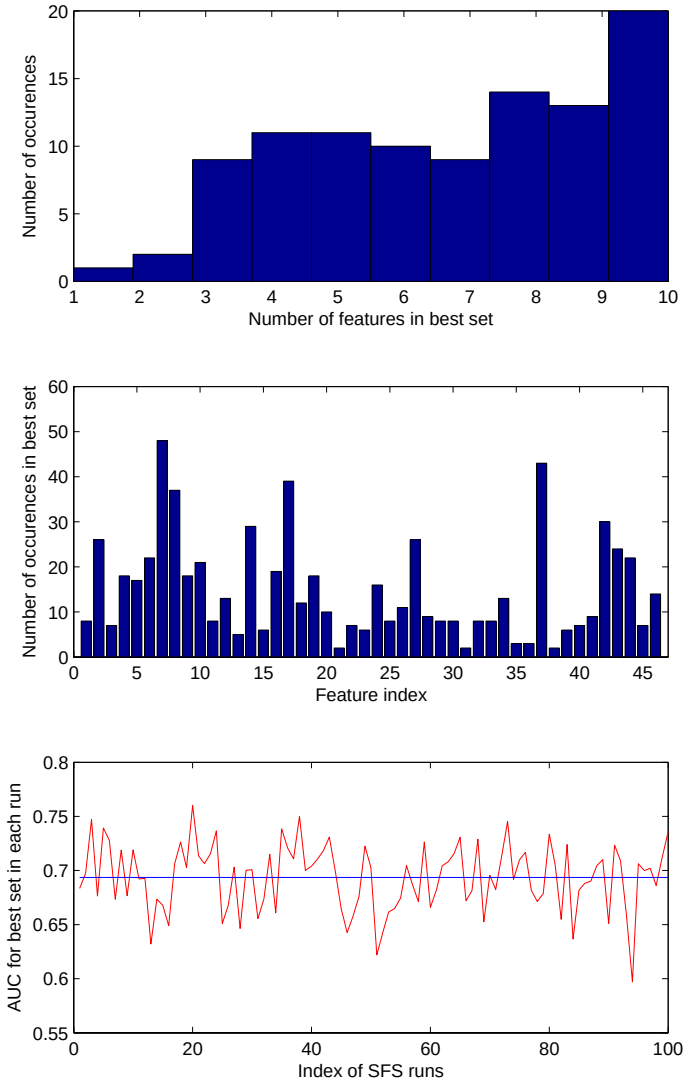



Figure 6.4: Feature selection statistics for jet features and stripiness features. Average $AUC = 0.69 \pm 0.03$

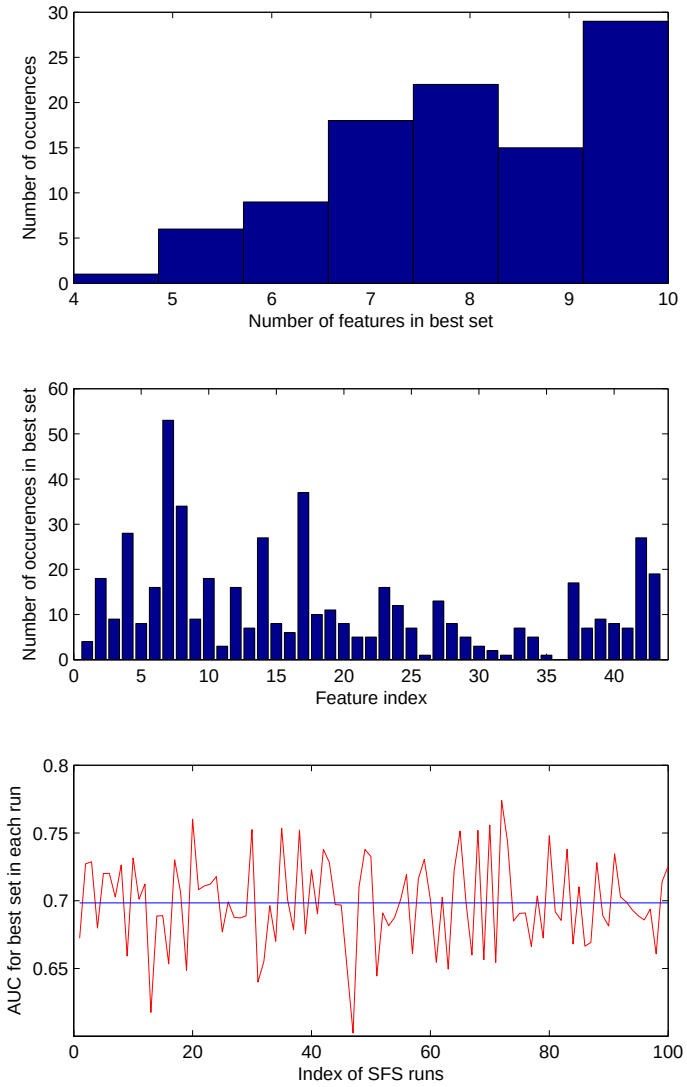



Figure 6.5: Feature selection statistics for jet features with stripiness features being forced in the initial selection. Average $AUC = 0.70 \pm 0.03$

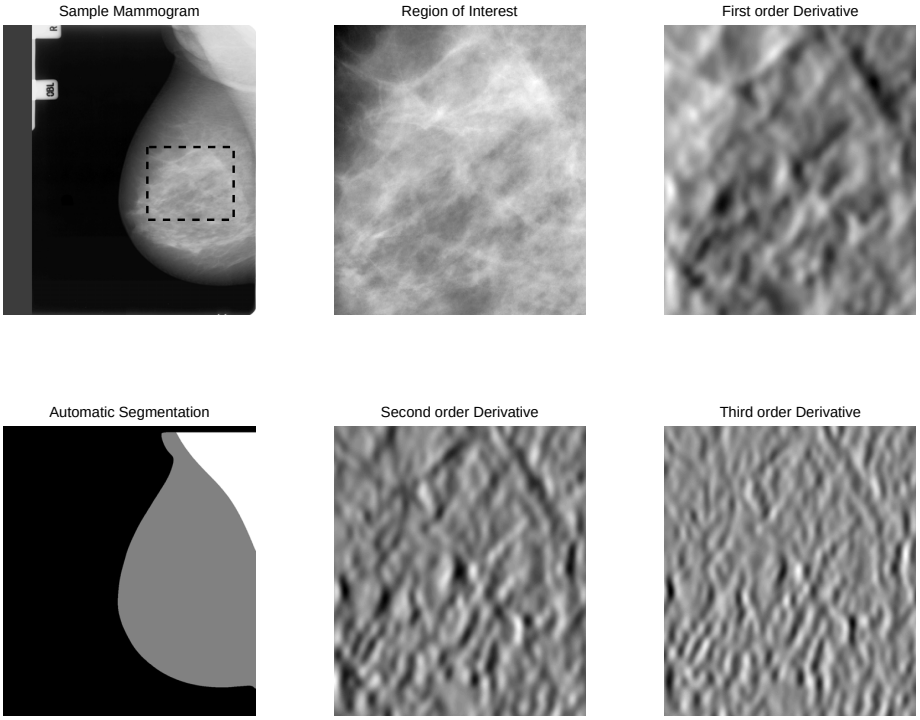
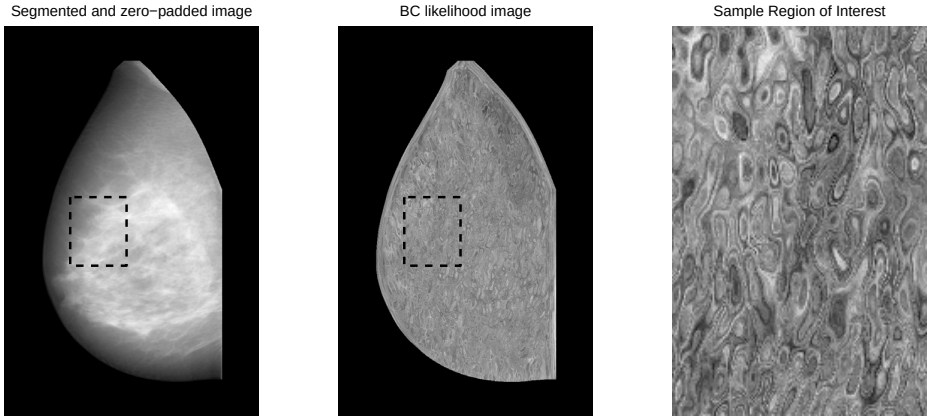



Figure 6.6: A sample mammogram from the investigated data, corresponding automatic segmentation, and a superimposed region of interest to illustrate the three horizontal derivatives (scale 1mm).

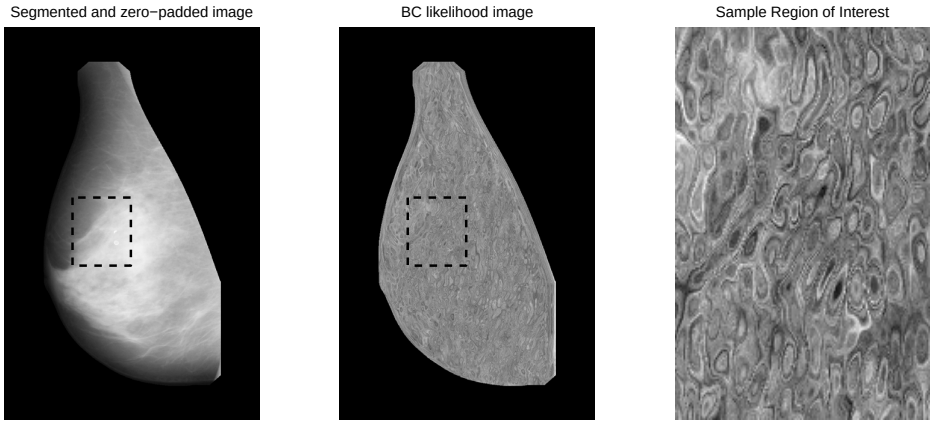
That the position features are important is supported by findings by Li et al., building on the work by Huo et al., showing a statistically significant decrease of performance as the location of the used region of interest (ROI) was varied from the central region immediately behind the nipple. Li et al. do not compare the results obtained using ROIs to using the whole breast area for feature estimation.

The stripiness features, shown in the previous two chapters to be indicative of HRT, appears to be only weakly related to risk. This is in line with findings by Boyd et al.[27] indicating that the effects of hormone therapy on mammographic density, and on breast cancer risk, are separate and not related causally.

To see how a mammogram and corresponding likelihood image actually look like we computed the likelihood images of two cases, using the feature set [7 17 27 37 42 43], and included them in Figure 6.7. Case (a) is from the same patient as displayed in Figure 6.6, who had a screen-detected cancer in the right breast four years later. The BIRADS score of the mammogram is 3 but the likelihood score is quite low, 48.9% compared to an average of $50.2 \pm 0.9$ for all the cases. Case (b) is an interval case also with




(a)




(b)

Figure 6.7: Two sample mammograms and corresponding likelihood images using features [7 17 27 37 42 43]. Case (a) has an average pixel probability of 48.9% and case (b) 52.6%

a BIRADS score of 3 but a higher likelihood score, 52.6%. Although it is difficult to relate the appearance of the likelihood images to the corresponding mammograms, it is clear from the zoomed regions of interest that there is some structure present.

To compare with results by Huo et al. the histogram skewness was computed for all the images. This was the single feature found most related to risk in [63]. One difference in implementation is that we compute the skewness of the entire breast region and Huo et al. use a smaller ROI. The skewness is one of the features found related to mammographic density by Boone et al. [42] and is related to the degree of symmetry of the histogram.

Huo et al. report an AUC of $0.82 \pm 0.04$ for discrimination of 15 BRCA1

/ BRCA2 mutation carriers versus 143 'low-risk' women. Classifying the images in the present study as cases or controls based on histogram skewness gave an AUC of 0.60. In comparison we on average get $0.70 \pm 0.03$ with the selected cancer features.

It is hard to do any further comparison. Where we investigate large sets of local features, Huo et al. use existing domain knowledge to construct and test a smaller number of global features. There is also a difference in evaluation; where we evaluate directly on case/control evidence Huo et al. use two ways of labelling high-risk patients: 1) Carriers of the BRCA1 / BRCA2 genetic mutation and 2) correlation with risk estimated by the Gail and Claus models.

This way of evaluating the results has some shortcomings. In the case of 1), the implications of the results are somewhat limited since only very few actual cancer cases carry the BRCA1 / BRCA2 mutation. One study [69] reporting a frequency of around 1% of 1220 investigated cases and another 2-3 % of 1628 cases [70]. Moreover, the evaluation was based on just 15 mutations carriers. In the case of 2), it is not possible to discover new risk factors in the examined image features, but just potential manifistations of known ones included in the Gail and Claus models. Also, breast density is not included as risk factor in these models.

Possible additions to our work, potentially leading to better separation of cancers and controls, include using a more sophisticated feature selection scheme and testing global image features including those proposed by Huo et al. [63].

In conclusion we have demonstrated the benefits of including feature selection to our proposed supervised framework in a medical setting. The proposed general methodology may be used to learn features for different diseases or treatments, potentially gaining insight to the biology behind different changes manifested in mammograms or in other medical image data. A feature selection experiment with 3-jet features and polynomial invariants up to third order showed that, generally, the 3-jet resulted in larger AUC when separating cases and controls. Additional results indicated that [7 17 27 37 42 43] (third order horizontal derivatives and horizontal and vertical distance to center of distance map) is a feature set indicative of breast cancer risk. Investigating scorings using these features showed higher odds ratios of incidence of breast cancer than using BI-RADS rating and automatic percentage density. These results are presented in Chapter 9.

# Part III

# Clinical Results

# Chapter 7

# Comparing the effects of orally and nasally dosed HRT on mammographic density and patterns[1]

## 7.1 Abstract

**Objectives:** To compare the impact of nasally and orally dosed estradiol on breast density. Secondarily, to investigate the utility of computer-based automated approaches to the assessment of breast density with reference to traditional methods.
**Methods:** Digitised images from two 2-year, randomised, placebo-controlled trials formed the basis of the present post hoc analysis. Active treatments were 1 mg estradiol continuously combined with 0.125 mg trimegestone (Oral HRT) or low-dose (150 or 300 $\mu$g estradiol) nasal estradiol cyclically combined with 200 mg micronised progesterone (Nasal HRT). The effects on breast density was assessed by a radiologist providing the BI-RADS score and the interactive threshold, and by computer-based approach providing the measure of stripiness and the HRT-effect specific measure of breast density.
**Results:** In the oral HRT trial, active treatment induced significant increase in breast density, which was consistent with all methods used (all $p < 0.05$). In contrast, none of the methods detected significant changes in women receiving nasal HRT. The sensitivity of automated methods to discriminate HRT- from placebo-treated women was equal or better than the methods performed by the radiologist.

[1]Larger parts of the clinical introduction and discussion were written by radiologist Paola Pettersen who also coauthored the clinical paper [17] on which this chapter is based.

**Conclusions:** The markedly different pharmacokinetic profile of nasal estrogen seems to be associated with better breast safety. Automated computer-based analysis of digitized mammograms provides a sensitive measure of changes in breast density induced by hormones, and could serve as useful tools in future clinical trials.

## 7.2 Introduction

Estrogen deficiency accompanying the menopause is a major pathophysiological mechanisms underlying accelerated bone turnover and bone loss after the menopause leading to increased risk of fragility fractures in the elderly age. Hormone replacement therapy (HRT) remains among the most rational approaches to prevent not only the short-term (e.g. vasomotor symptoms), but also the long-term consequences (osteoporosis) of the menopause [71]. Hormone therapy can be administered via different routes — orally, transdermally, nasally — all characterized by different dosing regimes and pharmacokinetic profiles. Whereas oral HRT for primary prevention is currently not recommended based on recent findings of the Women's Health Initiative trial [72], it remains to be clarified whether the adverse effects of conjugated estrogen (0.625 mg) plus medroxyprogesterone acetate (2.5 mg) are also applicable to lower doses, other combinations and administration routes.

Recent clinical development has led to the introduction of the nasal estradiol spray that has an entirely different pharmacokinetic profile compared with the traditional oral therapy [73]. After nasal administration, plasma estradiol levels rise rapidly and fall to 10% of the peak level within 2 hours [74], unlike both oral and transdermal administration, which both produce prolonged or plateau estrogen levels [75]. The efficacy of the nasal therapy in terms of controlling menopausal symptoms and preventing bone loss has been demonstrated by randomised clinical trials [73, 76], but the effects of the therapy on the breast tissue have not been assessed. Initial observations indicated that postmenopausal women taking nasal therapy report less frequent adverse events related to mastalgia than those taking oral treatment [77], suggesting that the pulsed therapy might be advantageous.

An important parameter when addressing breast safety is breast density, which was shown related to breast cancer risk in several studies [21]. Indeed, women with dense breasts may have a 2- to 6-fold increase risk of breast cancer [22].

The analysis of the breast density is normally performed by radiologists using the four categories of the Breast Imaging Report and Data System (BI-RADS), originally proposed by the American College of Radiology (ACR) [20]. Other methods requiring the interaction of a radiologist include the

interactive threshold measurement method that expresses dense areas as percentage of the total breast area [11]. This latter method carries relative advantages in terms of monitoring, because it provides a continuous measure of breast density. Whether these continuous measurements can be automated or improved needs the introduction of adequate computer-based methods and their testing in clinical trials side-by-side with traditional techniques.

In the present study, we set out to investigate 1) whether the pulsed dosing of HRT provides relative advantages in terms of breast safety compared with the oral administration route, and 2) whether the recently introduced automated methods can provide comparable or better approaches to the quantification of breast density than the currently widely used radiologist-assisted approaches.

## 7.3 Materials and Methods

### Subjects

The study population is from two previously published hormone trials assessing the efficacy and safety of oral or nasal estradiol combined with respectively continuous or cyclic progestin on postmenopausal loss [73, 49]. In both clinical trials, participants were between 40 and 65 years of age, postmenopausal for at least 1 year, and had a BMI equal or below 32 kg/m2 at study entry. Menopause was defined as consistent amenorrhoea for more than 12 months, or amenorrhoea for more than 6 months combined with serum level of estradiol below 0.16nmol/l and follicle stimulating hormone (FSH) level above 42 IU/l. All women were healthy with no clinical and laboratory evidence of systemic disease and had not been receiving any medication known to influence bone or lipid metabolism. They all had osteopenia, defined as a lumbar spine BMD between -1.0 and -2.5 SD of the premenopausal mean value. In the study utilizing the nasal spray route, the upper limit of the T-score was extended to +1 by amendment. Exclusion criteria ensured that none of the women had any contraindications for the use of HRT, nasal disease incompatible with nasal administration, or any suspicious breast lump detectable with bilateral mammography at baseline. In the original trials, mammography was a safety not an efficacy measure. A number of patients were randomised based on negative findings of mammography taken in other screening centre or hospital within 6 months before entry to the study. These subjects did not have baseline images for assessing 2-year changes of breast density, and hence were not included in the present analysis. In the nasal HRT trial, 267 subjects had mammography available for the two visits. In the oral HRT trial, only 76 subjects had mammography available for the two visits. All participants signed an ap-

proved informed consent to participation and both trials were carried out according to the Helsinki Declaration II and European Standards to Good Clinical Practice. The local ethical committees have approved the study protocols.

## Study designs and treatments

*The nasal HRT trial*
This was a 2-year, randomised, double blind, placebo-controlled clinical study that recruited patients at two study sites in Denmark (Ballerup and Aalborg). Patients were allocated randomly to receive treatment with either 150$\mu$g or 300$\mu$g estradiol (E2), which was administered once daily in the evening, or placebo for 2 years. Women with intact uterus also received 200 mg micronised progesterone, or progestin placebo (placebo patients), combined with the nasal spray in the last 14 days of each 28-day cycle.

*The oral HRT trial*
This was also a 2-year, multi-centre, double blind, placebo-controlled, randomized clinical trial investigating the efficacy and safety of 1 mg 17$\beta$-estradiol combined with 0.125 mg trimegestone for the prevention of postmenopausal osteoporosis. All subjects received a daily supplement of 500 mg calcium and 400IU of vitamin D.

## Breast density quantified by the radiologist

Mammography was obtained using a "Planmed Sophie" mammography X-ray unit. The right, medio-lateral image of each patient was processed for radiologist-assisted and automated image analysis. The images were digitized using a Vidar scanner providing an image resolution of 200 microns per pixel and 12-bit gray scales. On the digitized image, delineation of the breast boundary was done manually by the reading radiologist using points along the boundary connected with straight lines, resulting in a region of interest. When scoring the images the reading radiologist was blinded with respect to the labeling of the patients. The same radiologist made all readings.

**a) Categorical (BI-RADS)** — The BI-RADS categories are: 1) Entirely fatty; 2) Fatty with scattered fibroglandular tissue; 3) Heterogeneously dense; 4) Extremely dense. The reading radiologist assigned the mammograms to one of these categories based on visual inspection.

**b) Interactive threshold** — The reading radiologist determines an intensity threshold using a slider in a graphical user interface assisted visually by a display showing the region of dense tissue corresponding to the current slider position. The system is similar to the approach proposed by Byng et

al [11]. The density is defined as the ratio between segmented dense tissue and total area of breast tissue. Our implementation was made using the Matlab software (Mathworks, MA, USA).

### Breast density quantified by computer-based approach

**a) Stripiness** — In every pixel, at three different scales (1mm, 2mm, 4mm), we recorded the elongatedness of the local image structure. This is defined as the ratio between the difference and the sum of the eigen values of the local Hessian measured at these scales. This ultimately compares to the eccentricity of an ellipse approximating the local image. Each pixel is then assigned to one of the four different stripiness types, which are defined by the three measures of elongatedness. For a given image, the final stripiness measure is a specific weighted difference of the numbers of pixels in these classes [13]. Image analysis and calculations were performed using Matlab Software (Mathworks, MA, USA).
**b) HRT-effect specific breast density** — In every pixel, the elongatedness is measured at three different scales (1mm, 2mm, 4mm) as done for the stripiness measure. Subsequently, every pixel is compared to pixels from other mammograms, and it is recorded how many of the 100 most alike pixels found in the other mammograms are from patients treated with HRT, and how many are from patients treated with placebo. These counts act as votes for, respectively, HRT and placebo. The sum of votes from all pixels in the mammogram is counted and the percentile of votes for HRT is recorded as the final measure.

The measure has the same overall brightness and contrast invariance properties as the stripiness measure. It has been shown that this form of voting based on elongatedness at these scales can efficiently separate HRT and placebo groups [16]. Furthermore, it has been shown that it can likewise separate age groups if the voting is based on mammograms from different age groups. To avoid bias and overtraining issues patients were left out of the statistical analysis when their scores were computed. In other words, for each patient a statistical model is build on the N-1 remaining patients to provide a score.

### Statistical analysis

Data presented are expressed as mean $\pm$ standard error of the mean (SEM) unless otherwise indicated. Baseline characteristics of subjects in the different intervention groups were compared with Student's t-test for unpaired observations (oral HRT) or ANOVA (nasal HRT). Changes of breast density were assessed by Student's t-test for paired observations. The ability of the different methods to differentiate subjects, who received HRT for 2 years

Table 7.1: Characteristics of the study populations stratified by intervention groups.

| *Nasal spray route* | Placebo (n=98) | Low-dose (n=88) | High-dose (n=81) |
|---|---|---|---|
| Age (years) | $52.8 \pm 2.0$ | $52.6 \pm 1.6$ | $52.8 \pm 1.8$ |
| BMI ($kg/m^2$) | $24.9 \pm 3.6$ | $25.4 \pm 4.0$ | $25.4 \pm 4.1$ |
| Years since menopause | $2.3 \pm 1.4$ | $2.4 \pm 2.5$ | $2.0 \pm 1.2$ |
| Smoker, % | 25.5 | 27.2 | 30.9 |
| *Oral route* | Placebo (n=43) | E2 + TMG (n=46) | |
| Age (years) | $57.9 \pm 4.1$ | $58.2 \pm 3.8$ | |
| BMI ($kg/m^2$) | $24.9 \pm 3.3$ | $24.8 \pm 3.4$ | |
| Years since menopause | $7.8 \pm 5.0$ | $8.2 \pm 5.1$ | |
| Smoker, % | 20.9 | 17.4 | |

TMG: trimegestone, E2: estradiol

from those who have not, was also tested by unpaired t-test. Differences were considered statistically significant if $p < 0.05$.

## 7.4 Results

Table 7.1 outlines the baseline characteristics of subjects in the two clinical trials stratified in the different intervention groups. There were no statistically significant differences between subjects in terms of age, years since menopause, BMI, or BMD of the lumbar spine (different surrogate measures of endogenous estradiol).

The difference in smoking between groups within each trial is quite small (3-5% absolute). There is a somewhat larger percentage of smokers in the nasal trial (ca. 10% absolute) compared to the oral trial. The antiestrogenic effect of cigarette smoking causes smokers to have a slightly lower density on average than non-smokers [78]. However, this effect is not large enough to be observable in this study and therefore probably have no significant influence.

### Breast density at baseline

Using the BI-RADS method, the most frequent categorical finding at baseline examination was scattered density (47%) followed by heterogeneous density and fatty breasts (23% each) and extremely dense breasts (7%). The distributions of these categories as well as the mean values of breast density

Table 7.2: Different radiologist-assisted measures of breast density in the two clinical trials at baseline and after 2 years of hormone treatment

| | | NASAL HRT TRIAL | | | ORAL HRT TRIAL | |
|---|---|---|---|---|---|---|
| | | Placebo | Low | High | Placebo0 | E2 + TMG |
| BI-RADS | Baseline | 2.31 (0.10) | 2.09 (0.09) | 2.19 (0.08) | 2.30 (0.12) | 2.08 (0.16) |
| | End | 2.23 (0.09) | 2.06 (0.09) | 2.16 (0.09) | 2.28 (0.13) | 2.62* (0.15) |
| Threshold | Baseline | 0.21 (0.01) | 0.17 (0.01) | 0.19 (0.01) | 0.21 (0.02) | 0.20 (0.02) |
| | End | 0.20 (0.01) | 0.19 (0.02) | 0.20(0.02) | 0.21 (0.02) | 0.29*† (0.02) |

*, $p<0.001$ compared from baseline, †, $p<0.05$ compared with response in the placebo group, Threshold: Interactive Threshold.

obtained by the different methods were comparable between the different intervention groups in the two trials (all $p > 0.3$, Table 7.2).

### Changes of breast density (Radiologist)

In the clinical trial assessing the influence of oral HRT, both the BI-RADS categories and the interactive threshold techniques revealed significant increases in breast density from baseline in response to the 2-year treatment (both $p<0.001$, Table 7.2). Furthermore, results obtained by the interactive threshold method also indicated significant differences in breast density between placebo and hormone-treated patients at the follow-up visit ($p<0.05$). In contrast, in the nasal HRT trial, none of the methods indicated statistically significant increases in breast density from baseline, and there were no significant differences in breast density between placebo and HRT-treated patients at the end of the treatment period (Table 7.2).

### Changes in breast density (Computer-based methods)

Similar to the findings obtained with the radiologist-assisted methods, the computer-based methods also revealed statistically significant increases in breast density from baseline in the oral HRT and no changes in the nasal HRT trial (Table 7.3). In addition, the measure of stripiness and the HRT-effect specific measure of breast density also detected statistically significant differences between placebo and HRT-treated women at the end of the 2-year treatment period. In the nasal HRT trial, none of the methods captured statistically significant increases in the different measures of breast density from baseline, and hence no differences between responses of placebo and HRT-treated women.

Table 7.3: Different automated measures of breast density in the clinical trials at baseline and after 2 years of hormone therapy.

| | | NASAL HRT TRIAL | | | ORAL HRT TRIAL | |
|---|---|---|---|---|---|---|
| | | Placebo | Low | High | Placebo0 | E2 + TMG |
| Stripiness | Baseline | 0.133 (0.028) | 0.097 (0.020) | 0.103 (0.022) | 0.026 (0.005) | 0.027 (0.006) |
| | End | 0.094 (0.008) | 0.089 (0.018) | 0.092 (0.020) | 0.026 (0.005) | 0.043 (0.006)*† |
| Pix. Cl. | Baseline | 50.31 (0.08) | 50.14 (0.08) | 50.24 (0.07) | 49.86 (0.04) | 49.84 (0.04) |
| | End | 50.17 (0.08) | 50.10 (0.07) | 50.24 (0.08) | 49.80 (0.04) | 49.99 (0.05)*† |

*, $p<0.05$ compared from baseline, †, $p<0.05$ compared with response in the placebo group, Pix. Cl.: Pixel Classifier.

Table 7.4: Statistical significance of differences in breast density between HRT- and placebo-treated women at the end of the treatment period.

| | NASAL HRT TRIAL | ORAL HRT TRIAL |
|---|---|---|
| Threshold | 0.5 | 0.02 |
| Stripiness | 0.08 | 0.02 |
| Pix. Cl. | 0.95 | 0.001 |

*p*-values computed using two-sided, unpaired Student's t-tests. In the nasal HRT trial, the High-Dose Group represents the active treatment group.

### Comparison of the different techniques

As indicated by the results summarized in Table 7.4, there were notable differences in the sensitivity of the different methods for differentiating HRT-treated women from placebo-treated women at the end of the treatment period. The categorical BI-RADS scores were not able to discriminate hormone-treated subjects from placebo-treated ones in the oral HRT trial. In contrast, the continuous measure of breast density provided by the interactive threshold method was able to point out those who received HRT with reference to those who did not. The two computer-based techniques were both able to differentiate hormone- and placebo-treated patients with a sensitivity that was comparable to slightly better than that provided by the interactive threshold method (Table 7.4).

When applying the same methods on images collected in the nasal HRT trial, none of the methods indicated statistically significant differences between HRT- and placebo-treated women. There was however a tendency for marginal differences in terms of stripiness (p=0.08), which was in marked

contrast with the differences in terms of HRT-specific breast density measures (p=0.95).

## 7.5 Discussion

The two main findings of the present study were as follows: 1) oral administration of continuously combined estradiol plus trimegestone significantly increased breast density in postmenopausal women, which adverse effect were not visible when treating women with nasal estradiol cyclically combined with oral micronised progesterone, 2) the computer-based measures of breast density (stripiness and HRT-specific breast density) were comparable to somewhat more discriminative in the detection of differences in breast density in HRT- and placebo-treated postmenopausal women.

### Clinical aspects

Assessed by well-established and widely used techniques (BI-RADS and interactive threshold technique), daily oral dosing of 1 mg estradiol continuously combined with 0.125 mg trimegestone induced significant increases in breast density. These observations are in line with numerous earlier studies showing similar significant increases in breast density to treatment with continuously combined estrogen plus progestin [79, 80, 81, 82, 83, 84, 85]. Increases in breast density cannot be related to one particular combination or progestin per se, because similar observations were reported from trials testing combinations with norethisterone acetate, medroxyprogesterone acetate, or micronized progestin [79, 84, 86]. Since the effects of estradiol per se are minimal and rarely occur in patients [80], these observations point to the ethiopathogenic role of continuous progestin exposure of breast tissue.

In contrast, nasal administration of 150 or 300 $\mu$g estradiol cyclically combined with micronised progesterone did not seem to induce detectable changes in breast density. This observation confirms preliminary observations of Harma et al on a smaller group of subjects [10]. There are two main differences compared with the other regime, namely the mode of combining the progestin component and the administration route of estradiol. Arguing for the role of the progestin dosing regime, several side-by-side comparative studies pointed out that cyclic combination of the progestin or unopposed used of oral estrogen markedly reduces the incidence and magnitude of effect on breast density [79, 80, 81, 84, 85]. Thus, cyclic combination of the progestin component, and in particular when using bioidentical micronized progesterone, offers relative advantages for long-term use from the breast safety point of view.

However, nasal dosing of estradiol per se seems to provide an added value, as indicated by less frequently reports of mastalgia or other breast

discomfort in women treated with nasal estradiol cyclically combined with dydrogesterone compared with women receiving micronized oral estradiol with identical progestin dosing [87]. Recent clinical studies indicate that breast discomfort and breast density are closely related entities [88, 86]. The advantage of nasal estradiol may rest in its markedly different pharmacokinetics [89]. The main feature of nasal delivery of estradiol is a rapid peak followed by rapid normalisation of serum estradiol within 2 hours, implying a shorter pulsatile tissue exposure to estradiol compared with the relatively continuous exposure during oral HRT. Since we cannot exclude the possibility that interaction between the two sex steroids together responsible to the increases of breast density, minimizing the exposure to estradiol by nasal dosing and to progesterone by cyclic dosing seem logically beneficial for reducing potential adverse effects on breast tissue during HRT.

In summary, nasal estradiol cyclically combined with micronized progesterone appears to offer a gynecologically safer therapeutical choice for hormone replacement therapy; the advantages include less frequent withdrawal or breakthrough bleeding, less frequent mastalgia and an apparently negligible effect on breast density. These advantages may not only improve long-term compliance to therapy but may also culminate in less concerns for breast cancer, given the fact that increased breast density is often considered a surrogate measure of breast cancer risk in epidemiological studies [88].

### Methodological aspects

Changes of breast density were also assessed by two recently introduced image analysis programs that extract quantitative information independent of the subjective impressions of the radiologist. In general terms, both techniques were able to capture significant increases of breast density in postmenopausal women treated with oral HRT with sensitivity and accuracy comparable with traditional techniques.

There are though differences between the clustering approach (stripiness measure) and the effect-specific approach. The first is based in an intrinsic tissue subdivision not trained to recognize differences between labelled training data. This implies that the stripiness measure may capture structural changes that are not necessarily related to the hormone therapy. The second is an effect-specific method that was trained to demonstrate differences between labelled images (placebo vs. HRT).

The innovative element, recently introduced by Raundahl et al [16], is this ability to train the image analysis program to recognize changes of breast density attributable to specific effects. In the present context, when the program is trained by images from women who were treated or untreated with HRT in the past years, the software will automatically recognize and discriminate images showing differences in terms of likeli-

hood of the presence of increased breast density due to HRT. Importantly, the process of training does not introduce bias to the measurements, and the increases in breast density in the oral HRT trial are not merely the re-recognition of images, which the program was trained to see differences between.

The validity of negative findings regarding the changes of HRT-specific breast density is strongly supported by the facts that 1) the program was able to detect these effects in the oral trial and 2) the training session of the program was carried out using images from a completely other trial, further minimizing the introduction of methodological bias into the analysis of the nasal HRT trial.

Collectively, our multiple assessment of changes in breast density using assisted and non-assisted image analysis techniques provide consistent findings and thereby strong arguments for the apparent inert effects of pulsatile estrogen therapy on the hormone-sensitive parts of breast tissue. A further relative advantage of these computer-based techniques is independence of the qualifications and personal experience of the investigator, independence of image quality in terms of degree of x-ray exposition, and the fact that both of these techniques provide continuous measures of total breast density with statistical advantages.

## 7.6 Conclusion

In the present study, we showed that pulsatile hormone therapy via the nasal administration route may provide relative advantages in terms of breast safety compared with the apparent adverse effects of oral hormone therapy. Secondarily, we showed that automated computer-based analysis of digitised mammograms provides a sensitive measure of hormone-induced changes in breast density and could be a useful monitoring tool in future clinical trials assessing the safety of estrogen or hormone replacement therapies.

# Chapter 8

# Automatic scoring of mammographic patterns is more indicative of estrogen + progestogen treatment than breast density analyses performed by radiologist

There is some overlap between this chapter which focuses on the data from the oral HRT trial and the previous. The main updates of data are that we now have 90 images from the oral HRT trial[1] and an extra set of readings of follow-up mammograms. These new readings allow us to estimate the intra-observer variability of the two radiologist-assisted measures. In addition, a more thorough analysis and comparison of the cross-sectional separative ability of the BI-RADS, percentage density and pattern measures is conducted. Finally, an additional pattern scoring, which is trained on data from a breast cancer study, is included to investigate whether it is indicative of oral HRT.

## 8.1 Abstract

**Objectives:**: To investigate if computerized methodologies of quantising pattern changes are more indicative of estrogen + progestogen (E+P) induced changes than state-of-the-art scorings performed by radiologists. Secondarily to investigate whether a pattern-measure trained on breast cancer

[1] Retrieved 14 missing images from one of the centers

data is indicative of these changes.
**Methods:** Digitised images of completers of a two 2-year, randomised, placebo-controlled trial formed the base of the present post hoc analysis. Active treatment in the trial was 1 mg estradiol continuously combined with 0.125 mg trimegestone. Influence of the therapy on breast density was assessed with the following parameters: 1) categorical scores (BI-RADS), 2) computer-aided, interactive threshold method, 3) computerized HRT-effect specific scoring of breast patterns (HRT-Pattern), 4) computerized breast cancer specific scoring of breast patterns (BC-Pattern)
**Results:** E+P treatment induced highly significant increase of density/pattern scoring for all measures ($p<0.001$), except BC-Pattern. At follow-up, the categorical score and the interactive threshold showed significantly higher density in the treatment group (p=0.04, p=0.04); the HRT-Pattern scoring showed highly significant difference (p=0.001). HRT-Pattern scorings are moderately correlated with percentage dense area ($R^2$=0.68) and BI-RADS ($R^2$=0.51). After normalizing scores to identical mean and variance in placebo group at follow-up, the HRT-Pattern scoring was significantly more indicative than both the interactive threshold (p=0.04) and BI-RADS (p=0.04).
**Conclusions:** E+P treatment induced changes in the breast density as analysed by radiologist as well as in the pattern. These last changes can be automatically scored in a computerized fashion, and showed to be more sensitive than radiologist scorings. The HRT-Pattern score is mathematically invariant to density changes but showed correlations with the density measures in the present study. The BC-Pattern score was not indicative of HRT.

## 8.2 Introduction

The burden of estrogen deficiency in post menopausal women is the increased risk of fragility fractures due to an accelerated bone turnover, increasing bone loss [71]. Therefore hormone replacement therapy (HRT) is known to prevent postmenopausal osteoporosis and to reduce the incidence of fractures in osteoporotic women [90]. Nevertheless there is substantial scientific evidence that ovarian hormones, mainly estrogen plays a major role in the etiology of breast cancer [91]. This observation can be explained by the fact that estrogen replacement therapy can retard or even reverse the normal involutional process of the breast parenchyma [92, 93] leading to an increase of the density of the breasts.

Breast density is known as an important parameter for addressing breast safety. It estimates the proportion of fibroglandular tissue relative to the amount of fat. Moreover, breast density has been shown to relate to breast cancer risk in several large studies [22]. Indeed, these studies indicate that women with dense breasts have a 2- to 6-fold increased risk of breast cancer.

For this reason, breast density is often acknowledged as a surrogate marker of breast cancer risk and study parameter in diverse hormone-related prevention trials [94].

The analysis of the breast density is normally performed by radiologists using the four categories of the Breast Imaging Report and Data System (BI-RADS), originally proposed by the American College of Radiology [20]. Other methods still requiring the interaction of a radiologist include a threshold measurement that expresses dense areas as percentage of the total breast area [11]. This latter method carries relative advantages in terms of monitoring, because it provides a continuous measure of breast density.

In the present study, we sought to investigate whether automated scoring of the structural changes induced by HRT are more sensitive than the BI-RADS and the threshold analyses. In addition, we tested whether another pattern scoring, trained on data from a breast cancer study, was indicative of oral HRT.

## 8.3 Materials and methods

### Subjects

The study population is from a previously published hormone trial assessing the efficacy and safety of oral estradiol combined with continuous progestogen on postmenopausal bone loss [49]. Participants were between 40 and 65 years of age, postmenopausal for at least 1 year but less than 5 years, and had a BMI equal or below 32 kg/m2 at study entry. Menopause was defined as consistent amenorrhoea for more than 12 months, or amenorrhoea for more than 6 months combined with serum level of estradiol below 0.16nmol/l and follicle stimulating hormone (FSH) level above 42 IU/l. All women were healthy with no clinical and laboratory evidence of systemic disease and had not been receiving any medication known to influence bone or lipid metabolism. They all had osteopenia, defined as a lumbar spine BMD between $-1.0$ and $-2.5$ SD of the premenopausal mean value. Exclusion criteria ensured that none of the women had any contraindications for the use of HRT or any suspicious breast lump detectable with bilateral mammography at baseline.

In the original trial, mammography was a safety not an efficacy measure. A number of patients were randomised based on negative findings of mammography taken in other screening centres or hospitals within 6 months before entry to the study. These subjects did not have baseline images for assessing 2-year changes of breast density, and hence were not included in the present analysis. The original trial was a multi-centre trial and enrolled 360 patients. Of these all the patients from two centres were enrolled in the present study. Here 129 patients were enrolled, 94 (73%)

completed and 90 had both baseline and follow up mammograms. Of these 43 were from the placebo group and 47 from the treatment group. These subpopulations did not show any significant difference in age, years since menopause, BMI, % smokers compared to the original study population.

All participants signed an approved, informed consent to participation and both trials were carried out according to the Helsinki Declaration II and European Standards to Good Clinical Practice. The local ethical committees have approved the study protocols.

## 8.4 Study designs and treatments

The study was a 2-year, multi-centre, double blind, placebo-controlled, randomized clinical trial investigating the efficacy and safety of 1 mg 17$\beta$-estradiol combined with 0.125 mg trimegestone for the prevention of postmenopausal osteoporosis. All subjects received a daily supplement of 500 mg calcium and 400IU of vitamin D.

### Breast density quantified by the radiologist

Mammography was obtained using a "Planmed Sophie" mammography X-ray unit. The right, medio-lateral image of each patient was processed for radiologist-assisted and automated image analysis. The images were digitized using a Vidar scanner providing an image resolution of 200 microns per pixel and 12-bit gray scales. On the digitized image, delineation of the breast boundary was done manually by the reading radiologist using points along the boundary connected with straight lines, resulting in a region of interest. When scoring the images the reading radiologist was blinded with respect to the labeling of the images. Baseline and follow-up images were mixed and presented in random order. The same radiologist made all readings.

**a) Categorical (BI-RADS)** — The BI-RADS categories are: 1) Entirely fatty; 2) Fatty with scattered fibroglandular tissue; 3) Heterogeneously dense; 4) Extremely dense. The reading radiologist assigned the mammograms to one of these categories based on visual inspection.

**b) Interactive threshold** — The reading radiologist determines an intensity threshold using a slider in a graphical user interface assisted visually by a display showing the region of dense tissue corresponding to the current slider position. The system is similar to the approach proposed by Byng et al [11]. The density is defined as the ratio between segmented dense tissue and total area of breast tissue. Our implementation was made using the Matlab software (Mathworks, MA, USA).

## Breast density quantified by computer-based approach

**a) HRT indicative pattern scoring (HRT-Patterns)** — In every pixel, the elongatedness is measured at three different scales (1mm, 2mm, 4mm). This is defined as the ratio between the difference and the sum of the eigenvalues of the local Hessian measured at these scales [15]. Ultimately, this compares to the eccentricity of an ellipse approximating the local image structure. Subsequently, every pixel's eccentricity is compared to pixel eccentricities from the follow-up mammograms. It is recorded how many of the 100 most alike pixels found in the other mammograms are from patients treated with HRT, and how many are from patients treated with placebo. These counts act as votes for, respectively, HRT and placebo. The sum of votes from all pixels in the mammogram is counted and the percentile of votes for HRT is recorded as the final measure. Image analysis and calculations were performed using Matlab software (Mathworks, MA, USA).

The measure of elongatedness is invariant to global and certain local changes of image brightness and contrast [16]. It has been shown that elongatedness at these scales can efficiently separate HRT and placebo groups [16]. Furthermore, it was shown that it can likewise separate age groups if the voting is based on mammograms from different age groups. In computing the scores, a standard leave-one-out cross validation strategy [60] was employed, i.e., for each patient a statistical model is built on the $N-1$ remaining patients to provide a score.

**b) Breast cancer risk pattern scoring (BC-Patterns)** — The BC-Pattern scoring is similar to the HRT-Patterns only the voting was based on patients from a breast cancer study. Furthermore the features include position information and are based on filters describing the third order horizontal derivatives.

## Statistical analysis

Baseline characteristics of subjects in the different intervention groups were compared with Student's t-test for unpaired observations. Changes of breast density were assessed by Student's t-test for paired observations. The ability of the different methods to differentiate subjects, who received HRT for 2 years from those who have not, was tested by unpaired t-tests. All tests performed are two-sided and heteroscedastic unless otherwise indicated.

The comparison of different scoring techniques was performed by normalising the individual scores to having identical mean and variance in the placebo group at follow-up. One-sided, paired-sample Student's t-tests were performed on the normalised scores of the HRT group at follow-up to compare the effect of two scoring methodologies. A scoring methodology was considered significantly more effective if its mean of normalised HRT

Table 8.1: Correlations between normalised indicative scorings given as coefficient of correlation $R^2$ (95% confidence interval) for the placebo group, the treatment group, and the full study population. All correlations are significantly larger than zero ($p<0.001$).

| | Placebo | Treatment | All |
|---|---|---|---|
| BI-RADS vs. Thresholding | 0.90 (0.83–0.94) | 0.90 (0.82–0.94) | 0.90 (0.86–0.94) |
| BI-RADS vs. HRT-Pattern | 0.51 (0.27–0.70) | 0.64 (0.42–0.79) | 0.51 (0.27–0.70) |
| Thresholding vs. HRT-Pattern | 0.60 (0.37–0.75) | 0.71 (0.52–0.83) | 0.68 (0.54–0.77) |

scores was largest and $p<0.05$.

A further comparison was performed by assessing the intra-observer variability for the BI-RADS and percentage density methods. This was carried out by repeating the scoring of 90 random images after a period of several weeks. Intra-correlation, average change in scoring and number of exact agreements of assigned BI-RADS categories are reported.

## 8.5 Results

There were no statistically significant differences between subjects in terms of age, years since menopause, BMI, or BMD of the lumbar spine (different surrogate measures of endogenous estradiol). The HRT-Pattern score is mathematically invariant to density changes but showed correlations with the density measures in the present study (Table 8.1). The correlation between pattern and density was lower than between BI-RADS and percentage density.

### Changes of breast density (Radiologist)

In the clinical trial assessing the influence of the HRT, both the BI-RADS categories and the interactive threshold techniques revealed significant increases in breast density from baseline in response to the 2-year treatment (both $p<0.001$, Table 8.2). Furthermore, significant differences were also observed in breast density between placebo and hormone-treated patients at the follow-up visit ($p=0.04$ for both measures).

### Changes in breast pattern scoring (Computer-based method)

Similar to the findings obtained with the radiologist-assisted methods, the HRT-Pattern also revealed statistically significant increases in breast pat-

Table 8.2: Different scorings of mammograms in the two treatment groups at baseline and after 2 years of hormone treatment

| Measure | Time | Placebo | E2 + TMG |
|---|---|---|---|
| BC-Pattern | Baseline | $48.8 \pm 0.16$ | $48.7 \pm 0.20$ |
| | End | $48.8 \pm 0.22$ | $48.4 \pm 0.21$ |
| BI-RADS | Baseline | $2.28 \pm 0.13$ | $2.06 \pm 0.12$ |
| | End | $2.14 \pm 0.13$ | $2.45^{*\dagger} \pm 0.14$ |
| IA TH | Baseline | $0.21 \pm 0.03$ | $0.19 \pm 0.02$ |
| | End | $0.22 \pm 0.02$ | $0.29^{*\dagger} \pm 0.02$ |
| HRT-Pattern | Baseline | $49.87 \pm 0.05$ | $49.81 \pm 0.04$ |
| | End | $49.82 \pm 0.04$ | $50.00^{*\dagger} \pm 0.04$ |

Data shown are mean $\pm$ standard error of the mean, *: $p<0.001$ compared from baseline, †: $p<0.05$ compared with response in the placebo group. E2: Estradiol. TMG: trimegesterone. IA TH: Interactive threshold.

tern ($p<0.001$). In addition, the HRT-Pattern scoring also detected statistically highly significant differences between placebo and HRT-treated women at the end of the 2-year treatment period ($p=0.001$). There was no significant difference in BC-Pattern scoring between groups.

### Comparison of the different techniques

One-sided, paired-sample Student's t-tests of the normalised HRT scores at follow-up showed that HRT had a significantly higher effect on the computerized pattern scoring than on BI-RADS ($p=0.04$) and on percentage dense area ($p=0.04$).

Figure 8.1 illustrates the longitudinal change in placebo and HRT groups for the three methods. Tests for significance using unpaired Student's t-tests showed that the average change in scores were significantly higher ($p<0.001$) for the HRT group than for the placebo group measured with all three methods.

The intra-observer variability was larger for the BI-RADS scoring, $R^2$ = 0.79 (0.70–0.86), than for the interactive threshold scoring, $R^2$ = 0.95 (0.92–0.96). 68% of the BI-RADS scorings were in exact agreement and no cases disagreed more than one category. The average disagreement in BI-RADS score was 0.17 (0.05–2.8). The average disagreement in percentage area score was 0.05% (-1.1% – 1.2%).

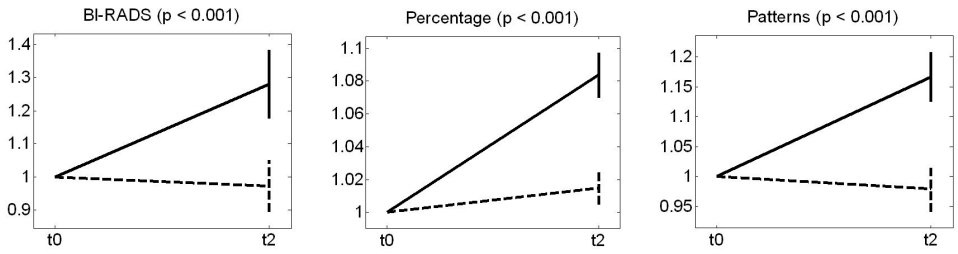



Figure 8.1: Figure 1. Illustration of longitudinal change of indicative measures in the placebo (dashed line) and HRT (solid line) groups. Base-line values are assigned the value 1 and the mean change in score is plotted. Vertical bars indicate STDOM of changes within the group.

## 8.6 Discussion

The main finding of the present study was that orally administrated estradiol combined with progestogen induced changes in patterns of mammograms which can be combined in a score which is significantly more indicative of HRT than both BI-RADS and percentage density at follow-up. There was no significant difference in patterns previously shown to be indicative of breast cancer risk. This does not necessarily translate to no increase in risk — more likely, there are either too few patients to show any effects or the type of general risk pattern reflected by the BC-Pattern is different from a potential, specific risk pattern induced by HRT.

Assessed by well-established and widely used techniques performed by radiologists (BI-RADS and interactive threshold technique), daily oral dosing of 1 mg estradiol continuously combined with 0.125 mg trimegestone induced significant increases in breast density. Numerous earlier studies revealed similar significant impact of oral HRT on breast density, especially when estrogen is combined continuously with progestin [79, 80, 81, 82, 83, 84, 85]. Increases in density seem to be independent of the type of progestin, given the similar observations related to norethisterone acetate, medroxyprogesterone acetate, and micronised progesterone [79, 84, 86]. Some of the studies investigating the safety of different dosing regimes in a side-by-side comparative study indicated that cyclic combination of the two components or unopposed estrogen therapy markedly reduces the impact of therapy on breast tissue [79, 80, 81, 84, 85]. Collectively, our findings are in line with the literature arguing for adverse impact of continuously combined estrogen plus progestin therapy on breast density.

### Methodological aspects

A promising perspective of the presented pattern recognition methodology is the potential of using other data to vote in the scoring process. In this way the method can be trained to pick up different effect-specific changes corresponding to e.g. other treatments versus placebo or diseased versus healthy patients.

A further relative advantage of the computer-based techniques is independence of the qualifications and personal experience of the investigator, independence of image quality in terms of degree of x-ray exposition, and the fact that both of these techniques provide continuous measures of total breast density with statistical advantages.

### Conclusion

In the present study, we showed that E+P therapy induced changes not only in the mammographic density, but also in the patterns. These subtle changes may be measured in a computerized fashion. Patterns relating to estradiol treatment were significantly more indicative than BI-RADS and percentage density at follow-up. There was no significant difference in BC-Pattern scoring between groups.

# Chapter 9

# Local pattern scoring of mammograms is a strong and independent predictor of breast cancer

## 9.1 Abstract

**Objectives**: To investigate to which degree local patterns indicate an elevation of breast cancer risk. Secondarily, to investigate to which degree these provide additional and more indicative information than breast density.
**Methods**: Digitised mammograms of 495 women were analysed. 245 of these women were diagnosed with breast cancer within 2-4 years, but radiological readings provided no evidence of cancer in radiographs. 250 women in age-matched control group without breast cancer diagnosis in the following 4 years. Relation to incidence of breast cancer for the following parameters was analysed: 1) categorical scores (BI-RADS), 2) computer-based percentage density method, 3) computer-based scoring of breast patterns.
**Results**: The mammographic scores were significantly higher for cases than controls (all $p < 0.001$). BI-RADS and percentage density are strongly correlated ($R^2$=0.41, 0.34–0.49) whereas the patterns are only weakly correlated with any of the two ($R^2$=0.11, 0.02–0.19) and ($R^2$=0.10, 0.01–0.18) for BI-RADS and percentage respectively. BI-RADS was not able to significantly separate cases and controls when adjusted for percentage density. Patterns could still separate cases and controls when adjusted for percentage density ($p < 0.001$) and vice versa ($p < 0.001$). The two measures carry mutually independent information and an aggregate measure combining the two gave consistent trend of increased separation of cancers and controls.

Investigating the parameters stratified corresponding to <5% dense versus >50% dense resulted in the following odds ratios: BI-RADS, 2.4 (1.4–4.1), Percentage density: 2.5 (1.5–4.2) and Patterns: 4.2 (2.4–7.2). The combined aggregate measure gave an odds ratio of 5.6 (3.2–9.8) which were significantly higher than for BI-RADS (p=0.03).
**Conclusions**: The odds ratio comparing high risk patterns to low risk patterns is up to 14.0 and always higher than or equal to both BI-RADS and percentage density scoring of mammographic density. The patterns are only weakly correlated with the density and seem to provide an independent indication of breast cancer risk. An aggregate measure combining this mutually independent information gave significantly higher odds ratio than using BI-RADS.

## 9.2 Introduction

An important parameter when addressing breast safety is breast density, which estimates the proportion of fibroglandular tissue relative to the amount of fat. Breast density has been shown to relate to breast cancer risk in several large studies [22]. Indeed, these studies indicate that women with dense breasts have a 2- to 6-fold increased risk of breast cancer.

The analysis of the breast density is normally performed by radiologists using the four categories of the Breast Imaging Report and Data System (BI-RADS), originally proposed by the American College of Radiology [20]. Other methods requiring the interaction of a radiologist include an interactive threshold measurement method that expresses dense areas as percentage of the total breast area [11]. This latter method carries relative advantages in terms of monitoring, because it provides a continuous measure of breast density.

General risk assessment does not take density into account. The Gail model [28] is a popular risk assessment tool and uses a woman's own personal medical history (number of previous breast biopsies and the presence of atypical hyperplasia in any previous breast biopsy specimen), her own reproductive history (number of reproductive years and age at the first birth of a child), and the history of breast cancer among her first-degree relatives (mother, sisters, daughters) to estimate her risk of developing invasive breast cancer over specific periods of time. Recent reports using breast density assessed by BI-RADS and continuous, planimetric measures found that the addition of breast density to the Gail model increased its ability to predict cases of breast cancer [29, 30].

Currently, however, the density is not used to assess risk in standard clinical screening procedures or included in general breast cancer risk assessment tools. A reason behind this is that, while breast density has become a well-established risk factor, the best way to measure, and indeed

what exactly to measure, is still a debated research topic [31].

Recent findings [44] have indicated that breast cancer risk is affected not only by the amount of mammographic density but also by the degree of heterogeneity of the breast pattern and, presumably, by other qualitative features captured by the Wolfe classification. This motivated us to include a measure based on local breast pattern appearance to investigate its potential as risk marker.

In the present study, we set out to investigate whether local patterns are indicative of an elevated breast cancer risk, and to which degree the information provided by these is independent from density markers.

## 9.3 Materials and methods

### Subjects

The study population of 495 women is from a previously published study [68] on the effect of recall rate in the Dutch biennial screening program . Participants were between 49 and 81 years of age. 245 women were subsequently diagnosed with breast cancer (123 interval and 122 screen-detected cancers). Mammograms were used from the screening 4 years prior to diagnosis for screen-detected cancers and 2-4 years prior to diagnosis for the interval cancers. In the control group mammograms of 250 women without breast cancer diagnosed in the subsequent 4 years were used. Cases and controls are age-matched apart from interval cases who are two years older due to the design of the original study.

### Breast density quantified by the radiologist

Categorical scorings of breast density by a trained radiologist were made as part of the original study are the radiologist classifications used in this investigation.

**Categorical (BI-RADS)** — The reading radiologist, blinded for cancer/case labelling, used four categories to analyse the breast density, giving a score from 1 to 4, where 1 denotes fatty, 2 scattered, 3 heterogeneously dense, and 4 extremely dense breast tissue. These categories were quantitatively defined as $< 5$ % dense, 5-25 % dense, 25-75 % dense, and $> 75$ % dense respectively.

### Breast density and patterns quantified by the computer

Prior to automated assessment the breast tissue was segmented automatically using techniques presented by Brady and Highnam [37] (breast boundary) and Karssemeijer [38] (pectoral muscle). Only the right mlo views

were analyzed in the computerized experiments.

**a) Automated percentage density** — An intensity threshold is separating the dense- and non-dense tissue is automatically determined. The system is similar to the approach proposed by Byng et al (6) only using a computerized approximation [12]. The density is defined as the ratio between segmented dense tissue and total area of breast tissue. Our implementation was made using Matlab software (Mathworks, MA, USA).

**b) Breast cancer risk pattern scoring (BC-patterns)** — In every pixel, a collection of multi-scale features are measured at three different scales (1mm, 2mm, 4mm, 8mm). Specifically the third order horizontal (relative to breast orientation) derivative was used. Furthermore, the position relative to the center [1] of the breast was recorded. Subsequently, these statistics for every pixel is compared to statistics of pixels from other mammograms, and it is recorded how many of the 100 most alike pixels found in the other mammograms are from cases and how many are from controls. These counts act as votes for, respectively, high risk and low risk. The sum of votes from all pixels in the mammogram is counted and the percentile of votes for risk is recorded as the final measure. To avoid bias and overtraining issues patients were left out of the statistical analysis when their scores were computed. In other words, for each patient a statistical model is build on the N-1 remaining patients to provide a score. Image analysis and calculations were performed using Matlab Software (Mathworks, MA, USA).

**c) HRT indicative pattern scoring (HRT-patterns)** — The HRT indicative pattern scoring is similar to the BC-pattern scoring only the voting was based on patients from an HRT trial receiving HRT or placebo treatment instead of cancers versus controls. Furthermore the features do not include position information and are based on derivative filters describing the local elongatedness. The HRT-patterns were shown to efficiently separate HRT and placebo groups in that study [16].

## Statistical analysis

Data presented are expressed as mean $\pm$ standard error of the mean (SEM) unless otherwise indicated. Group characteristics were compared with a heteroscedastic Student's t-test for unpaired observations. In addition, the different measures were compared to each other using a similar t-test on the cancer group normalising control group to zero mean and unit variance.

High risk and low risk thresholds of the measures were computed based on the F% fractile of the value of the measures. Odds ratios are computed

[1] The center of the breast was defined as the point with maximal distance to the closest breast boundary

Table 9.1: Age and measures according to patient stratification.

| | Control (n=250) | Interval (n=123) | Screen (n=122) | Cancers (n=245) |
|---|---|---|---|---|
| Age (years) | 57.3±0.4 | 59.8±0.6** | 57.7±0.5 | 58.8±0.4** |
| HRT Pattern | 50.61±0.02 | 50.67±0.02* | 50.64±0.02 | 50.65±0.01* |
| BI-RADS | 1.98±0.05 | 2.27±0.06** | 2.11±0.06 | 2.19±0.04** |
| Percentage | 13.0±0.7 | 18.6±1.0*** | 15.8±1.0* | 17.2±0.7*** |
| BC Pattern | 49.66±0.06 | 50.14±0.08*** | 50.16±0.08*** | 50.15±0.06*** |
| Aggregate | 50.29±0.07 | 51.07±0.10*** | 50.95±0.10*** | 51.01±0.07*** |

Asterisks indicate significant difference compared to control (* $p<0.05$, **$p<0.01$, ***$p<0.0001$).

for both high risk versus low risk and high risk versus rest of study population.

Odds ratios [95] and confidence intervals hereof are computed using the Mantel-Haenszel 95% confidence interval [96]. Odds ratio differences are tested by Tarone's adjustment [97] of the Breslow-Day test of heterogeneous odds ratios [98].

Correlation is computed as Pearson's coefficient of correlation, $R^2$, with associated p-value of being different from zero. Correlation coefficients are computed for data stratified into diagnosis groups to remove the influence of diagnosis.

The aggregate measure is computed using Fisher's linear discriminant analysis in a leave-one-out fashion [60]. This is to avoid any risk of bias or overfitting if the analysed mammogram is included in the statistical model. In a similar way, the high risk and low risk detection thresholds are computed using leave-one-out when computing the aggregate odds ratios.

## 9.4 Results

Table 9.1 shows a stratification of the different characteristics into the controls, interval cancers, screen-detected cancers, and all cancers. We remark again that the age of the interval cancer group is two years older than for controls and screen-detected cancers due to the design of the original study.

Table 9.2 shows the the normalized scores for the interval, screen-detected, and total cancers. Tables 9.3 and 9.4 show the odds ratios for high risk versus low risk patients using five percentile-based stratifications of the three most indicative measures, the percentage density, the BC-Pattern, and the aggregate score. In both tables the high risk patients are defined as having the F % highest scores. In Table 9.3 the low risk group is defined as having

Table 9.2: Normalised scorings according to patient stratification. All number are mean±SEM. Measurements are normalised so control group have zero mean and unit variance. The corresponding SEM's of controls are 0.06.

| | Interval (n=123) | Screen (n=122) | Cancers (n=245) |
|---|---|---|---|
| HRT Pattern | 0.24±0.08 | 0.12±0.08 | 0.18±0.06 |
| BI-RADS | 0.39±0.08 | 0.18±0.06 | 0.29±0.06 |
| Percentage | 0.53±0.09 | 0.27±0.10 | 0.40±0.07 |
| BC Pattern | 0.55±0.09 | 0.58±0.09*† | 0.57±0.07* |
| Aggregate | 0.76±0.10* | 0.64±0.10*† | 0.70±0.07*† |

*: p<0.05 when compared to BI-RADS, †: p<0.05 compared with area percentage.

Table 9.3: Odds ratios (95%CI) of high versus low risk patients defined as the highest F % of the population versus the lowest F %.

| F | 50% | 25% | 10% | 5% | 2% |
|---|---|---|---|---|---|
| Percentage | 1.9 (1.3-2.8) | 2.5 (1.5-4.1) | 2.4 (1.1-5.3) | 1.9 (0.7-5.1) | 1.5 (0.2-9.6) |
| BC Pattern | 2.6 (1.8-3.8) | 3.5 (2.0-6.0) | 9.8 (3.6-27)* | 8.1 (1.9-34)* | 14 (1.3-151)* |
| Aggregate | 2.8 (1.9-4.1) | 4.5 (2.6-8.0) | 4.9 (2.2-11) | 12 (3.2-42)* | 21 (2.0-217)* |

*: significantly different odds ratio compared to the same stratification by area percentage ($p < 0.05$ according to Tarone's adjustment of the Breslow-Day test of heterogeneous odds ratios).

the F % lowest scores and in Table 9.4 the low risk group consists of all the patients remaining after removing the high risk group.

Table 9.5 shows the statistical significance of differences between control and cancer for scores adjusted by influence of other measurements. Finally, Table 9.6 shows the correlation coefficients between different scores stratisfied in patient groups.

In addition, we did a small ROC analysis to compute the area under the curves (AUC) for the four measures. This resulted in the following areas. BI-RADS: AUC = 0.58, percentage density: AUC = 0.61, BC-pattern: AUC = 0.65, and aggregate: AUC = 0.68.

The number in each category of BI-RADS in ascending order is 106, 243, 144, and 2. High risk (BI-RADS >= 3) versus low risk (BI-RADS=1) gave an odds ratio (OR) of 2.4 (1.4–4.1). In comparison percentage density gave OR = 2.5 (1.5–4.2), BC-patterns: OR = 4.2 (2.4–7.2), and aggregate: OR = 5.6

Table 9.4: Odds ratios (95%CI) of high versus low risk patients defined as the highest F % of the population versus the remaining population.

| F | 50% | 25% | 10% | 5% | 2% |
|---|---|---|---|---|---|
| Percentage | 1.9 (1.3-2.8) | 2.3 (1.5-3.3) | 1.7 (1.0-2.9) | 2.1 (1.1-4.1) | 1.2 (0.4-3.7) |
| BC Pattern | 2.6 (1.8-3.8) | 2.5 (1.7-3.6) | 3.0 (1.8-4.9) | 2.1 (1.1-4.1) | 3.4 (1.3-8.9) |
| Aggregate | 2.8 (1.9-4.1) | 2.8 (1.9-4.1) | 3.3 (2.0-5.5) | 4.6 (2.4-8.8) | 5.2 (2.1-13)* |

*: significantly different odds ratio compared to the same stratification by area percentage ($p < 0.05$ according to Tarone's adjustment of the Breslow-Day test of heterogeneous odds ratios).

Table 9.5: Statistical significance of differences between control and cancer for scores (coloumn) adjusted by the influence of other scores (row).

| | BI-RADS | Percentage | BC-Pattern |
|---|---|---|---|
| BI-RADS | - | 0.1 | 0.008 |
| Percentage | 0.002 | - | 0.0002 |
| BC Pattern | $< 10^{-8}$ | $< 10^{-8}$ | - |

(3.2-9.8) for an equivalent stratification. The OR of the aggregate measure was significantly higher than for BI-RADS (p=0.03).

None of the measures were significantly different for patients with cancers occuring in right breast compared to those occuring in left. This demonstrates that the supervised measure, which is trained on right breasts only, was doing risk assessment and not some sort of early detection.

Table 9.6: Coefficient of correlation $R^2$ and its statistical significance between different scores in stratified patient groups.

| | Control | Interval | Screen |
|---|---|---|---|
| BI-RADS vs. Percentage | 0.43 ($< 10^{-11}$) | 0.41 ($< 10^{-5}$) | 0.33 (0.0002) |
| BI-RADS vs. BC-Pattern | 0.03 (0.64) | 0.09 (0.29) | 0.12 (0.18) |
| Percentage vs. BC-Pattern | -0.02 (0.68) | 0.18 (0.04) | 0.05 (0.54) |

## 9.5 Discussion

The BI-RADS and percentage density measures are more indicative of interval cancers than screen-detected cancers. This is probably because some cancers are overlooked due to dense tissue masking cancers. Since the masking effect is an increasing function of mammographic density [99] this will mean there are more patients with higher density among the inverval cancers. The BC-pattern does not exhibit this masking effect, it is equally indicative of screen-detected and interval cancers. In fact, the BC-Pattern is the only individual score effectively separating the screen-detected cases and the controls (Table 9.2).

BC-pattern scores are independent from BIRADS and percentage density since they do not correlate within groups (Table 9.6). BI-RADS and percentage density are correlated and once corrected for percentage density, BI-RADS carry no discriminative information (Table 9.5). BC-patterns are still very significantly indicative after adjustment for percentage density. So is percentage density after adjusting for BC-patterns, so a combination may be fruitful. This is reflected by the results showing higher AUC for the aggregate measure than any of the other measures.

The stratification we get when looking at the lowest BI-RADS category versus the two highest categories is similar to the one used by Torres-Mejia et al. [44] included in the meta-analysis of risk studies presented in [22]. They report and odds ratio of 3.49 (1.4–5.2) for a 0–5% versus $> 46\%$ stratification. This may be compared to our OR of 4.2 (2.4–7.2) for BC-Pattern and 5.6 (3.2-9.8) for the aggregate measure.

There are a number of reasons why the numbers are not directly comparable. The study reported by Torres-Mejia et al. is a cohort study where masking might raise the risks associated with mammographic density, since cancers missed in the first mammogram because of dense tissue would eventually be detected during subsequent follow-up. On the other hand they correct for the risk factors age and parity which again might lower our odds ratios slightly. The women in the cohort study were followed for 13-16 years and in comparison our odds ratio reflect a much shorter period of four years. All in all our findings indicate that the aggregate measure including both relative area of dense tissue and the BC-pattern risk score provides risk assessment which is superior to that of standard planimetric density.

The HRT-patterns does not correspond to BC-patterns. They do not separate cancers and controls and are uncorrelated with BC-patterns, percentage density, and BIRADS, p=0.70,p=0.20, and p=0.12 respectively when testing for significant correlation. They do give a significant indication of interval cancers, which is to be expected as the patterns were originally trained on data where treatment correlated with density, and the interval cancers have the highest density of the groups.

BC-patterns are very indicative of low risk patients but only slightly better than percentage density at high risk. This is indicated by the large difference in odds ratios for the BC-Pattern in Tables 9.3 and 9.4. This might indicate some biological difference between two types of dense tissue. High density without BC-patterns is not indicative of risk, but remains a useful risk marker in combination with an average or higher BC-pattern score.

If a framework to identify and quantify descriptions of the actual appearance of the patterns was developed this could be combined with knowledge from people approaching the topic from different angles, using other image modalities such as MRI or ultrasound and other fields of medical science, to get an integrated understanding of the underlying biological mechanisms.

A reliable risk measure derived automatically from a patient's mammogram has several possible practical implications. It can be used as additional safety measure in clinical trials. It can be used to develop a method for individually adjusting screening rates of women depending on their risk (of course combined with other known risk factors). During screening it might be used to set individual thresholds of suspiciousness so the recall rate becomes a dynamic function of risk.

## 9.6 Conclusion

Percentage density is more indicative of breast cancer risk than BI-RADS. Actually, BI-RADS proved redundant when adjusted for percentage density. BC-patterns are more indicative of risk than percentage density. The patterns are only weakly correlated with density and seem to provide an independent indication of breast cancer risk. An aggregate measure combining this mutually independent information gave significantly higher odds ratio than using BI-RADS alone.

# Part IV

# Closure

# Chapter 10

# Conclusion

This chapter contains a summary of the thesis, a discussion of its findings, several of their implications, and possible future work. Finally a short conclusion is given.

## 10.1 Summary

There are numerous studies showing that Hormone Replacement Therapy (HRT) increases density [6, 7, 8, 9] and that women with high breast density have a higher risk of breast cancer [22]. The causality of these effects is unclear and this motivated us to analyse images from HRT trials and from a breast cancer study to investigate and develop measures of density, patterns, and risk.

In Chapter 3, initial results using a threshold measure of mammographic density added more evidence for HRT induced mammographic density increase. After two years of treatment the average density of the HRT population was significantly higher than that of the placebo group ($p < 0.001$). Furthermore it was shown that the benefit of having two views compared to one is the same whether it is two projective views or using both left and right breast.

Subsequent experiments, presented in Chapter 4, showed that unsupervised clustering of mammograms, based on a specific quotient of Hessian eigenvalues at three scales, can be used to differentiate between patients receiving HRT and patients receiving placebo. The proposed mammographic pattern score was able to quantify the effect of HRT as structural changes in the breast tissue. To our knowledge the Hessian eigenvalues have not been used in connection with density in any previous work.

In Chapter 5 we introduced a supervised methodology based on a general statistical machine learning framework, using the same Hessian-based features, capable of differentiating different effect specific structural changes of the breast tissue. We have showed that changes in mammographic ap-

pearance caused by aging and HRT can be accessed as structural patterns ignoring the actual brightness of the images.

In Chapter 6 we presented a framework for incorporating feature selection in our supervised methodology. This framework was applied to a set of data from the Dutch national breast cancer screening program. The presented results demonstrated the ability and potential of including feature selection to improve and specialize measures. We found local mammographic features, mainly describing the structure around the vertical axis and the position in the breast, which were indicative of women developing breast cancer.

Chapter 7 showed that pulsatile hormone therapy via the nasal administration route provided relative advantages in terms of breast safety compared with the apparent adverse effects of oral hormone therapy. Secondarily, we showed that automated computer-based analysis of digitised mammograms provides a sensitive measure of hormone-induced changes in breast density and could be a useful monitoring tool in future clinical trials assessing the safety of estrogen or hormone replacement therapies.

In Chapter 8 we showed that estradiol induces changes not only of the mammographic density, but also in the patterns. These subtle changes may be measured in a computerized fashion. Patterns relating to estradiol treatment (HRT-Patterns) are more indicative than BI-RADS and percentage density at follow-up. Percentage density was not significantly more indicative of HRT than BI-RADS, but had significantly lower intra-observer variability. There was no significant difference in patterns shown indicative of breast cancer (BC-Patterns) risk between groups.

In Chapter 9 we demonstrated that percentage density is more indicative of breast cancer risk than BI-RADS. In fact, BI-RADS proved redundant when adjusted for percentage density. BC-patterns were more indicative of risk than percentage density. BC-patterns and percentage density carry mutually independent information and an aggregate measure combining the two scores gave superior odds ratios. HRT-Patterns was not found to be indicative of risk.

## 10.2 Discussion

The overall problem faced, methodologically, was deriving a measure capable of separating patients who are represented by images each consisting of millions of pixels. Common to the approaches we presented is that they rely on properties which can be derived locally, and not on global measures such as e.g. shape of breast boundary, symmetry measures or topological properties of segmented fibroglandular tissue. In short, each local area in the image gives a vote on the label of the patient associated with the image. For the most basic method, the automatic thresholding, this voting is

based on the smallest local area possible, namely individual pixels, and the voting is simply the pixel intensity. To get from the individual votes to a global score for the image the proportion of pixels with an intensity-vote above a specific intensity threshold is calculated.

In the case of unsupervised clustering each pixel is represented by features at three scales (1, 2, and 4mm) describing local elongatedness. Based on this representation the pixel gives a categorical vote of 2, $-1$ or 0 depending on its distance to four cluster centers representing areas in the mammograms with different structural properties.

The supervised approach again uses features to represent the pixels but base the voting directly on distances to sampled features from patients from the classes to be separated. This, combined with feature selection, gives the most adaptable and specific measure capable of generalizing to different separation tasks.

A possible addition to the methodology is a robust preprocessing step for normalization of mammograms ensuring equivalence of features and derived classifications. One could investigate combining the proposed supervised framework with something like the $h_{int}$ representation proposed by Brady and Highnam [37]. Although there will be problems for uncalibrated images Highnam et al presented a method [100] for estimating $h_{int}$ for uncalibrated images. This problem of calibration will disappear gradually with the advance of Full Field Digital Mammography (FFDM).

Another, larger project which would be interesting to investigate is including temporal and localized information using some form of registration technique. In this way the methodology could be used to identify and track pattern changes in individual patients instead of detecting more overall differences in cross-sectional data which was the focus of our investigation. A pilot study using differences of registrated images to track individual regeneration and involution of dense tissue was presented in [101] and, combined with our proposed framework, this could be taken as a point of departure for such investigations. Using this kind of tracking on views of both left and right breast could be of potential use in detecting asymmetric development of densities and patterns.

### 10.2.1 Clinical perspectives

Medical biomarkers can be related to various effects, such as a specific treatment or disease, or different demographic, reproductive, or anthropometric characteristics. Breast density is such a marker and is related to several effects (the major being age, menopausal status, HRT use and risk of breast cancer). We have described an automated approach of quantifying both 'traditional' density by adaptive thresholding and effect-specific pattern-measures using unsupervised and supervised pattern recognition techniques. A general advantage of more sensitive quantifications of biomark-

ers is a corresponding decrease in the number of participants needed in a clinical trial.

A reliable risk measure derived automatically from a patient's mammogram has several possible practical implications. It can be used as additional safety measure in clinical trials. It can be used to develop a method for individually adjusting screening rates of women depending on their risk (of course combined with other known risk factors). During screening it might be used to set individual thresholds of suspiciousness so the recall rate becomes a dynamic function of risk.

It could be very interesting to apply the proposed methodology to more data, learning and comparing various patterns. In this way additional knowledge of disease inter-relations could be gained. If a framework to identify and quantify descriptions of the actual appearance of the patterns was developed this could be combined with knowledge from people approaching the topic from different angles, using other image modalities such as MRI or ultrasound and other fields of medical science, to get an integrated understanding of the underlying biological mechanisms.

The gradual advance of FFDM makes the use of computerized assistance more natural both in terms of data collection and integration of digital methods. This does not hold only for risk assessment but also for any other computer-aided technique, such as detection or segmentation. We expect better reliability of image features and higher effective resolution enabling the investigation of smaller structures at a finer scale than currently possible when using digitizers as an intermediate step.

## 10.3 Conclusion

Numerous previous studies have shown that breast density is a strong predictor of breast cancer in women [22]. We showed that incorporating information about local patterns effectively doubled the association to risk of standard breast density measures in a data set of 245 cases and 250 controls. The odds ratio of breast cancer for an aggregate measure combining percentage density and patterns was 5.6 (3.2–9.8) compared to 2.4 (1.4–4.1) for BI-RADS and 2.5 (1.5–4.2) for percentage density. Patterns alone gave an odds ratio of 4.2 (2.4–7.2).

The final methodological framework is not quantifying density in the classical sense of estimating the ammount of fibroglandular tissue. Instead, it is a general and adaptable pattern recognition tool which can be trained to detect changes in available data. Because of the supervised machine learning approach employed, the method can be adapted to the detection of various mammographic changes. This makes it possible to do investigations of potential inter-relations of risk and different types treatment through cross-validation on data from clinical studies.

As demonstrated in the second part of this thesis, the approach is capable of building classifiers indicative of HRT, aging, and risk of breast cancer. Our findings indicate that effect-specific measures of HRT are not indicative of risk and vice versa. This is in line with findings by Boyd et al. [27] indicating that the effects of hormone therapy on mammographic density, and on breast cancer risk, are separate and not related causally. Still, we cannot explain the connection of HRT and risk, but having isolated independent markers of risk and HRT brings us one step closer to the elusive link between the associations of high mammographic density and breast cancer risk and of increased mammographic density and HRT use.

Generally, our findings using standard density measures are in agreement with published literature. Density was shown to increase under oral HRT and to be higher for patients developing cancer than matched controls. Nasal administration of HRT was shown to have no signficant influence on the breast density or patterns which probably translates to advantages in terms of breast safety. The continuous nature of the threshold density combined with the visual feedback from the scoring tool ment that percentage density overall performed better than the categorical BI-RADS in predicting risk and detecting changes caused by HRT.

# Bibliography


[1] B. M. ter Haar Romeny, L. M. J. Florack, A. H. Salden, and M. A. Viergever, "Higher order differential structure of images," *Image and Vision Computing*, vol. 12, no. 6, pp. 317–325, July/August 1994.

[2] D. M. Parking, P. Pisani, and J. Ferlay, "Global cancer statistics," *CA A Cancer Journal for Clinicians*, vol. 49, no. 1, pp. 33–64, jan-feb 1999.

[3] D. B. Kopans, *Breast Imaging*, 2nd ed. Lippincott - Raven, 1998.

[4] J. Couzin and M. Enserink, "More questions about hormone replacement," *Science*, vol. 298, p. 942, November 2002.

[5] R. Chlebowski, S. Hendrix, R. Langer, M. Stefanick, M. Gass, D. Lane, R. Rodabough, M. Gilligan, M. Cyr, C. Thomson *et al.*, "Influence of estrogen plus progestin on breast cancer and mammography in healthy postmenopausal women the women's health initiative randomized trial," *JAMA*, vol. 289, no. 24, pp. 3243–3253, 2003.

[6] G. A. Greendale, B. A. Reboussin, S. Slone, C. Wasilauskas, M. C. Pike, and G. Ursin, "Postmenopausal hormone therapy and change in mammographic density," *Journal of the National Cancer Institute*, January 2003.

[7] N. Colacurci, F. Fornaro, P. D. Franciscis, M. Palermo, and W. del Vecchio, "Effects of different types of hormone replacement therapy on mammographic density." *Maturitas*, vol. 40, November 2001.

[8] G. A. Greendale, B. A. Reboussin, A. Sie, H. R. Singh, L. K. Olson, O. Gatewood, L. W. Bassett, C. Wasilauskas, T. Bush, and E. Barrett-Connor, "Effects of estrogen and estrogen-progestin on mammographic parenchymal density. postmenopausal estrogen/progestin interventions (pepi) investigators." *Annals of Internal Medicine*, vol. 2, p. 262 269, February 1999.

[9] F. Sendag, M. T. Cosan, S. Ozsener, K. Oztekin, O. Bilgin, I. Bilgen, and A. Memis, "Mammographic density changes during different

postmenopausal hormone replacement therapies," *Annals of Internal Medicine*, vol. 76, September 2001.

[10] M. Harma, A. Öztürk, and M. Harma, "The effect of intranasal 17 $\beta$-estradiol on mammo graphic breast density," *Maturitas*, vol. 52, no. 2, pp. 165–166, 2005.

[11] J. W. Byng, N. F. Boyd, E. Fishell, R. A. Jong, and M. J. Yaffe, "The quantitative analysis of mammographic densities," *Physics in Medicine and Biology*, vol. 39, p. 162938, 1994.

[12] J. Raundahl, M. Nielsen, O. F. Olsen, and Y. Z. Bagger, "Effect of projective viewpoint in detecting temporal density changes," in *Medical Imaging 2004: Image Processing*, J. M. Fitzpatrick and M. Sonka, Eds., vol. 5370, 2004, pp. 85–92.

[13] J. Raundahl, M. Loog, and M. Nielsen, "Mammographic density measured as changes in tissue structure caused by HRT," in *Medical Imaging 2006: Image Processing*, J. M. Reinhardt and J. P. W. Pluim, Eds., vol. 6144, 2006, pp. 141–148.

[14] J. Raundahl, M. Loog, P. Pettersen, and M. Nielsen, "Evaluation of four mammographic density measures on HRT data," in *Medical Imaging 2007: Image Processing*, vol. 6512, 2007.

[15] J. Raundahl, M. Loog, and M. Nielsen, "Understanding hessian-based density scoring," in *Digital Mammography*, vol. 4046. Springer Berlin / Heidelberg, September 2006, pp. 447–452.

[16] J. Raundahl, M. Loog, P. Pettersen, and M. Nielsen, "Quantifying effect-specific mammographic density," in *Medical Image Computing and Computer-Assisted Intervention MICCAI 2007*, vol. 4792. Springer Berlin / Heidelberg, October 2007, pp. 580–587.

[17] P. Pettersen, J. Raundahl, M. Loog, M. Nielsen, L. B. Tank, and C. Christiansen, "Parallel assessment of the impact of different hrts on breast density by radiologist- and computer-based analyses of mammograms," *Climacteric*, February 2008, in press. The two first authors contributed equally to the paper.

[18] S. Shapiro, "Evidence on screening for breast cancer from a randomized trial," *Cancer*, vol. 39, no. 2772-2778, 1977.

[19] J. N. Wolfe, "Risk for breast cancer development determined by mammographic parenchymal pattern," *Cancer*, vol. 37, no. 5, pp. 2486–2498, 1976.

[20] A. C. of Radiology, *Illustrated Breast Imaging Reporting and Data System*, 3rd ed. American College of Radiology, 1998.

[21] J. A. Harvey and V. E. Bovbjerg, "Quantitative assessment of mammographic breast density: Relationship with breast cancer risk," *Radiology*, vol. 230, no. 1, pp. 29–41, January 2004.

[22] N. F. Boyd, J. M. Rommens, K. Vogt, V. Lee, J. L. Hopper, M. J. Yaffe, and A. D. Paterson, "Mammographic breast density as an intermediate phenotype for breast cancer," *The Lancet Oncology*, vol. 5, pp. 798–808, 2005.

[23] J. Brisson, B. Brisson, G. Cot, E. Maunsell, S. Brub, and J. Robert, "Tamoxifen and mammographic breast densities," *Cancer Epidemiology Biomarkers and Prevention*, vol. 9, no. 9, pp. 911–915, September 2000.

[24] M. Freedman, J. S. Martin, J. O'Gorman, S. Eckert, M. E. Lippman, S.-C. B. Lo, E. L. Walls, and J. Zeng, "Digitized mammography: a clinical trial of postmenopausal women randomly assigned to receive raloxifene, estrogen, or placebo," *Journal of the National Cancer Institute*, vol. 93, no. 1, pp. 51–56, January 2001.

[25] C. H. van Gils, J. H. C. L. Hendriks, R. Holland, N. Karssemeijer, J. D. M. Otten, H. Straatman, and A. L. M. Verbeek, "Changes in mammographic breast density and concomitant changes in breast cancer risk," *European Journal of Cancer Prevention*, vol. 8, pp. 509–515, 1999.

[26] K. Kerlikowske, L. Ichikawa, D. L. . Miglioretti, D. S. . M. . Buist, P. M. . Vacek, R. Smith-Bindman, B. Yankaskas, P. A. Carney, and R. Ballard-Barbash, "Longitudinal measurement of clinical mammographic breast density to improve estimation of breast cancer risk," *Journal of the National Cancer Institute*, vol. 99, no. 5, pp. 386–395, March 2007.

[27] N. Boyd, L. Martin, Q. Li, L. Sun, A. Chiarelli, G. Hislop, M. Yaffe, and S. Minkin, "Mammographic density as a surrogate marker for the effects of hormone therapy on risk of breast cancer," *Cancer Epidemiology Biomarkers & Prevention*, vol. 15, no. 5, p. 961, 2006.

[28] M. H. Gail, L. A. Brinton, D. P. Byar, D. K. Corle, S. B. Green, C. Schairer, and J. J. Mulvihill, "Projecting individualized probabilities of developing breast cancer for white females who are being examined annually," *Journal of the National Cancer Institute*, vol. 81, no. 24, pp. 1879–86, December 1989.

[29] M. Palomares, J. Machia, C. Lehman, J. Daling, and A. McTiernan, "Mammographic density correlation with gail model breast cancer risk estimates and component risk factors," *Cancer Epidemiology Biomarkers & Prevention*, vol. 15, no. 7, p. 1324, 2006.

[30] J. Tice, S. Cummings, E. Ziv, and K. Kerlikowske, "Mammographic breast density and the gail model for breast cancer risk prediction in a screening population," *Breast Cancer Research and Treatment*, vol. 94, no. 2, pp. 115–122, 2005.

[31] J. Couzin, "Detecting a hidden breast cancer risk," *Science*, vol. 309, pp. 1664–1666, September 2005.

[32] J. W. Byng, N. F. Boyd, E. Fishell, R. A. Jong, and M. J. Yaffe, "Automated analysis of mammographic densities," *Physics in Medicine and Biology*, vol. 41, pp. 909–923, 1996.

[33] W. Berg, C. Campassi, P. Langenberg, and M. Sexton, "Breast imaging reporting and data system inter-and intraobserver variability in feature analysis and final assessment," *American Journal of Roentgenology*, vol. 174, no. 6, pp. 1769–1777, 2000.

[34] C. Zhou, H.-P. Chan, N. Petrick, M. A. Helvie, M. M. Goodsitt, B. Sahiner, and L. M. Hadjiiski, "Computerized image analysis: Estimation of breast density on mammograms," *Medical physics*, vol. 28, no. 6, pp. 1056–1069, June 2001.

[35] R. Sivaramakrishna, N. A. Obuchowski, W. A. Chilcote, and K. A. Powell, "Automatic segmentaion of mammographic density," *Academic Radiology*, vol. 8, pp. 250–256, 2001.

[36] J. Shepherd, L. Herve, J. Landau, B. Fan, K. Kerlikowske, and S. Cummings, "Novel use of single x-ray absorptiometry for measuring breast density," *Technology in Cancer Research and Treatment*, vol. 4, no. 2, pp. 173–82, 2005.

[37] R. Highnam and M. Brady, *Mammographic Image Analysis*, M. A. Viergever, Ed. Kluwer Academic Publishers, 1999.

[38] N. Karssemeijer, "Automated classification of parenchymal patterns in mammograms," *Physics in Medicine and Biology*, vol. 43, pp. 365–378, 1998.

[39] X. Wang, W. Good, B. Chapman, Y. Chang, W. Poller, T. Chang, and L. Hardesty, "Automated assessment of the composition of breast tissue revealed on tissue-thickness-corrected mammography," *American Journal of Roentgenology*, vol. 180, pp. 257–62, January 2003.

[40] A. Oliver, J. Freixenet, R. Marti, and R. Zwiggelaar, "A comparison of breast tissue classification," *MICCAI*, pp. 872–9, 2006.

[41] M. Jeffreys, R. Warren, R. Highnam, and G. Davey Smith, "Initial experiences of using an automated volumetric measure of breast density: the standard mammogram form," *British Journal of Radiology*, vol. 79, no. 941, p. 378, 2006.

[42] J. M. Boone, K. K. Lindfors, C. S. Beatty, and J. A. Seibert, "A breast density index for digital mammograms based on radiologists' ranking," *Journal of Digital Imaging*, vol. 11, no. 3, pp. 101–115, August 1998.

[43] P. Miller and S. Astley, "Classification of breast tissue by texture analysis," *Image and Vision Computing*, vol. 10, no. 5, pp. 277–282, 1992.

[44] G. Torres-Mejia, B. D. Stavola, D. S. Allen, J. J. Perez-Gavilan, J. M. Ferreira, I. S. Fentiman, and I. dos Santos Silva, "Mammographic features and subsequent risk of breast cancer: A comparison of qualitative and quantitative evaluations in the guernsey prospective studies," *Cancer Epidemiology Biomarkers and Prevention*, vol. 14, no. 5, pp. 1052–59, May 2005.

[45] S. Petroudi and M. Brady, "Breast density segmentation using texture," in *International Workshop on Digital Mammography*, S. M. Astley, M. Brady, C. Rose, and R. Zwiggelaar, Eds. Springer, 2006, pp. 609–615.

[46] E. Warner, G. Lockwood, D. Trichler, and et al, "The risk of breast cancer associated with mammographic parenchymal patterns: a meta-analysis of the published literature to examine the effect of method of classification," *Cancer Detection & Prevention*, vol. 16, pp. 67–72, 1992.

[47] J. W. Byng, N. F. Boyd, L. Little, G. Lockwood, E. Fishell, R. A. Jong, and M. J. Yaffe, "Symmetry of projection in the quantitative analysis of mammographic images," *European Journal of Cancer Prevention*, vol. 5, pp. 319–327, 1996.

[48] J. Kittler and J. Illingworth, "Minimum error thresholding," *Pattern Recognition*, vol. 19, no. 1, pp. 41–47, 1986.

[49] L. Warming, P. Ravn, D. Spielman, P. Delmas, and C. Christiansen, "Trimegestone in a low-dose, continuous-combined hormone therapy regimen prevents bone loss in osteopenic postmenopausal women." *Menopause*, vol. 11, no. 3, pp. 337–342, May-June 2004.

[50] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. Wiley-Interscience, 2001.

[51] M. A. Maloof, S. V. Beiden, and R. F. Wagner, "Analysis of competing classifiers using components of variance of roc accuracy measures," *CS*, vol. 1, 2002.

[52] H. Rickard, G. Tourassi, and A. Elmaghraby, "Unsupervised tissue segmentation in screening mammograms for automated breast density assessment," *Proceedings of SPIE*, vol. 5370, pp. 75–84, 2004.

[53] A. Frangi, W. Niessen, K. Vincken, and M. Viergeve, "Multiscale vessel enhancement filtering," in *Medical Image Computing and Computer-Assisted Intervention - MICCAI'98*. Springer, 1998, pp. 130–137.

[54] J. J. Koenderink, "The structure of images," *Biological cybernetics*, vol. 50, no. 5, pp. 363–370, 1984.

[55] B. ter Haar Romeny, *Front-End Vision and Multi-Scale Image Analysis*. Kluwer Academic Publisher, 2003.

[56] F. van der Heiden, R. Duin, D. de Ridder, and D. Tax, *Classification, Parameter Estimation, State Estimation: An Engineering Approach Using MatLab*. Wiley, New York, 2004.

[57] R. Duin, P. Juszczak, P. Paclik, E. Pekalska, D. de Ridder, and D. Tax, "Prtools4, a matlab toolbox for pattern recognition," *Delft University of Technology*, 2004.

[58] X.-B. Pan, M. Brady, R. Highnam, and J. Declerck, "The use of multi-scale monogenic signal on structure orientation identification and segmentation," in *International Workshop on Digital Mammography*, S. M. Astley, M. Brady, C. Rose, and R. Zwiggelaar, Eds. Springer, 2006, pp. 601–608.

[59] C. Tromans and M. Brady, "An alternative approach to measuring volumetric mammographic breast density," in *International Workshop on Digital Mammography*, S. M. Astley, M. Brady, C. Rose, and R. Zwiggelaar, Eds. Springer, 2006, pp. 26–33.

[60] A. K. Jain, R. P. W. Duin, and J. Mao, "Statistical pattern recognition: A review," *IEEE Tr. on PAMI*, vol. 22, no. 1, pp. 4–37, 2000.

[61] S. Arya, D. Mount, N. Netanyahu, R. Silverman, and A. Wu, "An optimal algorithm for approximate nearest neighbor searching in fixed dimensions," in *ACM-SIAM. Discrete Algorithms*, no. 5, 1994.

[62] M. Freedman, J. S. Martin, J. O'Gorman, S. Eckert, M. E. Lippman, S. B. Lo, E. L. Walls, and J. Zeng, "Digitized mammography: a clinical trial of postmenopausal women randomly assigned to receive raloxifene, estrogen, or placebo," *Journal of the National Cancer Institute*, vol. 93, no. 1, pp. 51–56, January 2001.

[63] Z. Huo, M. Giger, D. Wolverton, W. Zhong, S. Cumming, and O. Olopade, "Computerized analysis of mammographic parenchymal patterns for breast cancer risk assessment: Feature selection," *Medical Physics*, vol. 27, p. 4, 2000.

[64] E. Claus, N. Risch, and W. Thompson, "Autosomal dominant inheritance of early-onset breast cancer. implications for risk prediction." *Cancer*, vol. 73, no. 3, pp. 643–51, 1994.

[65] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *The Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[66] A. Whitney, "A direct method of nonparametric measurement selection," in *IEEE Trans. Comput.*, vol. 20, 1971, pp. 1100–1103.

[67] J. Koenderink and A. van Doorn, "Representation of local geometry in the visual system," *Biological Cybernetics*, vol. 55, no. 6, pp. 367–375, 1987.

[68] J. D. M. Otten, N. Karssemeijer, J. H. C. L. Hendriks, J. H. Groenewoud, J. Fracheboud, A. L. M. Verbeek, H. J. de Koning, and R. Holland, "Effect of recall rate on earlier screen detection of breast cancers based on the dutch performance indicators," *Journal of the National Cancer Institute*, vol. 97, no. 10, pp. 748–754, May 2005.

[69] A. B. C. S. Group, "Prevalence and penetrance of brca1 and brca2 mutations in a population-based series of breast cancer cases," *British Journal of Cancer*, vol. 83, pp. 1301–1308, 2000.

[70] K. Malone, J. Daling, D. Doody, L. Hsu, L. Bernstein, R. Coates, P. Marchbanks, M. Simon, J. McDonald, S. Norman *et al.*, "Prevalence and predictors of brca1 and brca2 mutations in a population-based study of breast cancer in white and black american women ages 35 to 64 years," *Cancer Research*, vol. 66, no. 16, p. 8297, 2006.

[71] M. Davidson, K. Maki, P. Marx, A. Maki, M. Cyrowski, N. Nanavati, and J. Arce, "Effects of continuous estrogen and estrogen-progestin replacement regimens on cardiovascular risk markers in postmenopausal women," *Archives of Internal Medicine*, vol. 160, no. 21, pp. 3315–3325, 2000.

[72] J. Lacey Jr, P. Mink, J. Lubin, M. Sherman, R. Troisi, P. Hartge *et al.*, "Risks and benefits of estrogen plus progestin in healthy postmenopausal women: Principal results from the womens health initiative randomized controlled trial," *Obstetrics & Gynecology*, vol. 101, pp. 194–197, 2003.

[73] T. Nielsen, P. Ravn, Y. Bagger, L. Warming, and C. Christiansen, "Pulsed estrogen therapy in prevention of postmenopausal osteoporosis. a 2-year randomized, double blind, placebo-controlled study," *Osteoporosis International*, vol. 15, no. 2, pp. 168–174, 2004.

[74] J. Devissaguet, N. Brion, O. Lhote, and P. Deloffre, "Pulsed estrogen therapy: pharmacokinetics of intranasal 17-beta-estradiol (s21400) in postmenopausal women and comparison with oral and transdermal formulations." *Eur J Drug Metab Pharmacokinet*, vol. 24, no. 3, pp. 265–71, 1999.

[75] R. Scott, B. Ross, C. Anderson, and D. Archer, "Pharmacokinetics of percutaneous estradiol: a crossover study using a gel and a transdermal system in comparison with oral micronized estradiol," *acogjnl*, vol. 77, no. 5, pp. 758–764, 1991.

[76] J. Studd, B. Pornel, I. Marton, J. Bringer, C. Varin, Y. Tsouderos, and C. Christiansen, "Efficacy and acceptability of intranasal 17 [beta]-oestradiol for menopausal symptoms: Randomised dose-response study." *Obstetrical & Gynecological Survey*, vol. 54, no. 10, pp. 645–646, 1999.

[77] N. Yesildaglar, S. Erkaya, D. Uygur, K. Göl, B. Bingöl, and Z. Günenç, "Efficacy of pulsed estrogen therapy in relatively younger patients with surgically induced menopause," *Human Reproduction*, vol. 19, no. 1, pp. 210–213, 2004.

[78] Y. Bremnes, G. Ursin, N. Bjurstam, and I. T. Gram, "Different measures of smoking exposure and mammographic density in postmenopausal," *Breast Cancer Research*, vol. 9, no. 5, October 2007, epub ahead of print.

[79] Y. Bremnes, G. Ursin, N. Bjurstam, E. Lund, and I. Gram, "Different types of postmenopausal hormone therapy and mammographic density in norwegian women," *Int J Cancer*, vol. 120, no. 4, pp. 880–4, 2007.

[80] N. Topal, S. Ayhan, U. Topal, and T. Bilgin, "Effects of hormone replacement therapy regimens on mammographic breast density: The role of progestins," *Journal of Obstetrics and Gynaecology Research*, vol. 32, no. 3, pp. 305–308, 2006.

[81] S. Pongsatha, M. Muttarak, S. Chaovisitseree, S. Luewan, and A. Panpanit, "Mammographic changes related to different types of hormonal therapies." *J Med Assoc Thai*, vol. 89, no. 2, pp. 123–9, 2006.

[82] D. Marchesoni, L. Driul, A. Ianni, G. Fabiani, M. Della Martina, C. Zuiani, and M. Bazzocchi, "Postmenopausal hormone therapy and mammographic breast density," *Maturitas*, vol. 53, no. 1, pp. 59–64, 2006.

[83] P. Conner, G. Svane, E. Azavedo, G. Söderqvist, K. Carlström, T. Gräser, F. Walter, and B. von Schoultz, "Mammographic breast density, hormones, and growth factors during continuous combined hormone therapy," *Fertility and Sterility*, vol. 81, no. 6, pp. 1617–1623, 2004.

[84] F. Sendag, M. Cosan Terek, S. Ozsener, K. Oztekin, O. Bilgin, I. Bilgen, and A. Memis, "Mammographic density changes during different postmenopausal hormone replacement therapies." *Fertil Steril*, vol. 76, no. 3, pp. 445–50, 2001.

[85] E. Lundstrom, B. Wilczek, Z. von Palffy, G. Soderqvist, and B. von Schoultz, "Mammographic breast density during hormone replacement therapy: differences according to treatment," *Am J Obstet Gynecol*, vol. 181, no. 2, pp. 348–352, 1999.

[86] N. Bülbül, S. Özden, and V. Dayicioglu, "Effects of hormone replacement therapy on mammographic findings," *Archives of Gynecology and Obstetrics*, vol. 268, no. 1, pp. 5–8, 2003.

[87] L. Mattsson, C. Christiansen, J. Colau, S. Palacios, P. Kenemans, C. Bergeron, O. Chevallier, T. Von Holst, and K. Gangar, "Clinical equivalence of intranasal and oral 17beta-estradiol for postmenopausal symptoms," *Am J Obstet Gynecol*, vol. 182, no. 3, pp. 545–52, 2000.

[88] C. Crandall, A. Karlamangla, M. Huang, G. Ursin, M. Guan, and G. Greendale, "Association of new-onset breast discomfort with an increase in mammographic density during hormone therapy," *Archives of Internal Medicine*, vol. 166, no. 15, p. 1578, 2006.

[89] A. Genazzani and F. Bernardi, "Estrogen effects on neuroendocrine function: the new challenge of pulsed therapy." *Climacteric*, vol. 5, no. 2, pp. 50–6, 2002.

[90] T. DJ and B.-S. SE, "Hormone replacement therapy and prevention of nonvertebral fractures: a meta-analyses of randomized trials," *Journal of the American Medical Association*, vol. 285, pp. 2891–2898, 2001.

[91] M. Cotterchio, N. Kreiger, B. Theis, M. Sloan, and S. Bahl, "Hormonal factors and the risk of breast cancer according to estrogen-and progesterone-receptor subgroup 1," *Cancer Epidemiology Biomarkers & Prevention*, vol. 12, no. 10, pp. 1053–1060, 2003.

[92] L. Titus-Ernstoff, A. Tosteson, C. Kasales, J. Weiss, M. Goodrich, E. Hatch, and P. Carney, "Breast cancer risk factors in relation to breast density (united states)," *Cancer Causes and Control*, vol. 17, no. 10, pp. 1281–1290, 2006.

[93] F. van Duijnhoven, P. Peeters, R. Warren, S. Bingham, P. van Noord, E. Monninkhof, D. Grobbee, and C. van Gils, "Postmenopausal hormone therapy and changes in mammographic density," *Journal of Clinical Oncology*, vol. 25, no. 11, p. 1323, 2007.

[94] S. Gapstur, P. Lopez, L. Colangelo, J. Wolfman, L. Van Horn, and R. Hendrick, "Associations of breast cancer risk factors with breast density in hispanic women 1," *Cancer Epidemiology Biomarkers & Prevention*, vol. 12, no. 10, pp. 1074–1080, 2003.

[95] J. Bland, "Statistics notes: The odds ratio," *BMJ*, vol. 320, no. 7247, pp. 1468–1468, 2000.

[96] N. Mantel and W. Haenszel, "Statistical aspects of the analysis of data from retrospective studies," *Journal of the National Cancer Institute*, vol. 22, no. 4, pp. 719–48, 1959.

[97] R. Tarone, "On heterogeneity tests based on efficient scores," *Biometrika*, vol. 72, no. 1, p. 91, 1985.

[98] N. Breslow and N. Day, *Statistical methods in cancer research. Vol. 1, The analysis of case-control studies*. International Agency for Research on C6ancer, 1980.

[99] C. van Gils, J. Otten, A. Verbeek, and J. Hendriks, "Mammographic breast density and risk of breast cancer: Masking bias or causality?" *European Journal of Epidemiology*, vol. 14, no. 4, pp. 315–320, 1998.

[100] R. Highnam, J. Brady, and B. Shepstone, "Estimation of compressed breast thickness during mammography," *British Journal of Radiology*, vol. 71, no. 846, p. 646, 1998.

[101] S. Petroudi, K. Marias, and M. Brady, "Evaluation of effects of hrt on breast density," *Lecture Notes in Computer Science*, vol. 4046, p. 39, 2006.