Quantifying Effect-Specific Mammographic Density

Jakob Raundahl¹, Marco Loog^{1,2}, Paola Pettersen³, and Mads Nielsen^{1,2}

¹ University of Copenhagen, Department of Computer Science, Denmark

² Nordic Bioscience, Herlev, Denmark

³ Center for Clinical and Basic Research, Ballerup, Denmark

Abstract. A methodology is introduced for the automated assessment of structural changes of breast tissue in mammograms. It employs a generic machine learning framework and provides objective breast density measures quantifying the specific biological effects of interest. In several illustrative experiments on data from a clinical trial, it is shown that the proposed method can quantify effects caused by hormone replacement therapy (HRT) at least as good as standard methods. Most interestingly, the separation of subpopulations using our approach is considerably better than the best alternative, which is interactive. Moreover, the automated method is capable of detecting age effects where standard methodologies completely fail.

1 Introduction

Women with high mammographic density appear to have a four to six fold increase in breast cancer risk, as indicated by numerous studies [1,2,3,4]. Since this makes breast density a surrogate measure of risk for developing cancer in the breast, a sensitive measure of changes in this density would be most valuable and could aid as a diagnostic tool for the analysis of mammograms. A related issue concerns the influence of post-menopausal hormone replacement therapy, which induces an increase in mammographic density [5,6,7]. During hormone dosing, such a sensitive measure would provide a reliable indicator of the gynecological safety of a given treatment modality.

The classical way to measure breast density is to use a categorical score, such as either Wolfe patterns [1] or the breast imaging reporting and data system (BI-RADS) [8]. These measures have been constructed to explain different kinds of biological and mammographic effects. The main purpose of Wolfe patterns, as an example, is to indicate breast cancer risk and BI-RADS mainly indicates masking effect.

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 [9,10,11,12,13]. 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 method is invariant to affine intensity changes.

We mean to convince the reader of the fact that density changes can indeed be 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 2, introduces the medical study that produced the images used in this investigation. Subsequently, Section 3 describes the two standard methods and the new supervised method in detail. Section 4 contains a description of the experimental setup and results. Section 5 consists of discussions and conclusion.

2 Materials

Since HRT has been shown to increase mammographic density[5,6,7], images from a placebo controlled HRT study may be used to evaluate density measures by their ability to separate the HRT and placebo populations. Furthermore, aging effects can be detected by stratifying the baseline patients according to age.

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 β -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, at least 1 year postmenopausal with a body mass index less than or equal to 32 kg/m².

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, using 10 points along the boundary connected with straight lines, resulting in a decagon region of interest. 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 [14]. We denote the patient groups P0, P2, H0, and H2 for placebo and treatment at t_0 and t_2 respectively.

3 Methods

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

3.1 BI-RADS

Breast imaging reporting and data system (BI-RADS) is the four category scheme proposed by the American College of Radiology [8]. 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 [15].

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 [16] and has been used in several clinical trials [15]. The density is defined as the ratio between segmented dense tissue and total area of breast tissue.

3.3 The Supervised Approach

Our density measure is derived by training a pixel classifier on subsets of images from the available data. These subsets are chosen to represent the different changes in density 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 the two different sets 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. [17]), but we do not necessarily expect benefit from these. An example of a mammogram with corresponding pixel probability maps is shown in Fig. 1. Below follows a more

Fig. 1. Mammogram from the data set (a); pixel classification result using the classifiers HRTC, HRTL and AGE respectively (b), (c), (d)

precise description of the features and a description of the various subgroups used to train the classifiers.

Features. A specific three dimensional feature space is used since a previous study found these features to be associated with breast density [18]. These 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 $G * \partial I = I * \partial G$ [19]. 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 ϵ is a small positive number ($\epsilon = 10^{-5}$) to avoid numerical stability problems in the accidental and non-generic 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 classifier ($k = 100$) is trained using this data to separate pixels from the two classes. The use of this powerful, non-parametric classifier is justified by the large number of pixels and low dimensionality of the feature space.

Due to the limited amount of patients, the data is not split up into a 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. This technique is similar to leave-one-out [17], but is modified to leave-two-out since leaving a sample out from one class creates a bias in the training set toward the other. This effect is unwanted, especially when dealing with a small number of samples. In the training phase, feature vectors from 10,000 randomly selected pixels within the breast region in each image are used, which is sufficient considering that our feature space is only three dimensional.

4 Experimental Setup and Results

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 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. 2 the density changes are shown using the three different training strategies together with the BI-RADS scores and the

Table 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

Fig. 2. Relative 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_2 and of the entire baseline population at t_0 .

Fig. 3. AGE density as a function of age in tertiles in the baseline population including the standard deviation of the mean for each tertile

percentage density. The actual values of the different scores are immaterial and are not displayed. The figure allows for a qualitative comparison of the methods by showing the relative progressions of the HRT and placebo groups combined with standard deviations of means.

Fig. 3 shows that the differences between P0 and P2 indicated by the AGE classifier is indeed an age related effect and not a general difference in image appearance at t_0 and t_2 . The baseline population is again stratified into three age groups and the AGE measures show an increasing trend with increasing age. The values for the first and last tertile are significantly different ($p = 0.015$).

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 randomization. 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.

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. Ultimately, it may be possible to train the system to accurately quantify breast cancer risk, providing better risk assessment than the standard density measures. Clearly, we may need additional features to detect these different effects, but our general and effective framework can readily cope with such extensions.

References

1. Wolfe, J.N.: Risk for breast cancer development determined by mammographic parenchymal pattern. *Cancer* 37(5), 2486–2498 (1976)
2. Boyd, N.F., O’Sullivan, B., Campbell, J.E., et al.: Mammographic signs as risk factors for breast cancer. *British Journal of Cancer* 45, 185–193 (1982)
3. Boyd, N.F., Byng, J.W., Jong, R.A., Fishell, E.K., Little, L.E., Miller, A.B., Lockwood, G.A., Trichler, D.L., Yaffe, M.J.: Quantitative classification of mammographic densities and breast cancer risk: Results from the canadian national breast screening study. *Academic Radiology* 87(9), 670–675 (1995)

4. van Gils, C.H., Hendriks, J.H.C.L., Holland, R., Karssemeijer, N., Otten, J.D.M., Straatman, H., Verbeek, A.L.M.: Changes in mammographic breast density and concomitant changes in breast cancer risk. *European Journal of Cancer Prevention* 8, 509–515 (1999)
5. Greendale, G.A., Reboussin, B.A., Slone, S., Wasilauskas, C., Pike, M.C., Ursin, G.: Postmenopausal hormone therapy and change in mammographic density. *Journal of the National Cancer Institute* (January 2003)
6. Colacurci, N., Fornaro, F., Franciscis, P.D., Palermo, M., del Vecchio, W.: Effects of different types of hormone replacement therapy on mammographic density. *Maturitas* 40 (November 2001)
7. Sendag, F., Cosan, M.T., Ozsener, S., Oztekin, K., Bilgin, O., Bilgen, I., Memis, A.: Mammographic density changes during different postmenopausal hormone replacement therapies. *Annals of Internal Medicine* 76 (September 2001)
8. American College of Radiology: Illustrated Breast Imaging Reporting and Data System, 3rd edn. American College of Radiology (1998)
9. Boone, J.M., Lindfors, K.K., Beatty, C.S., Seibert, J.A.: A breast density index for digital mammograms based on radiologists' ranking. *Journal of Digital Imaging* 11(3), 101–115 (1998)
10. Karssemeijer, N.: Automated classification of parenchymal patterns in mammograms. *Physics in Medicine and Biology* 43, 365–378 (1998)
11. Byng, J.W., Boyd, N.F., Fishell, E., Jong, R.A., Yaffe, M.J.: Automated analysis of mammographic densities. *Physics in Medicine and Biology* 41, 909–923 (1996)
12. Tromans, C., Brady, M.: An alternative approach to measuring volumetric mammographic breast density. In: Astley, S.M., Brady, M., Rose, C., Zwiggelaar, R. (eds.) *International Workshop on Digital Mammography*, pp. 26–33. Springer, Heidelberg (2006)
13. Petroudi, S., Brady, M.: Breast density segmentation using texture. In: Astley, S.M., Brady, M., Rose, C., Zwiggelaar, R. (eds.) *International Workshop on Digital Mammography*, pp. 609–615. Springer, Heidelberg (2006)
14. Byng, J.W., Boyd, N.F., Little, L., Lockwood, G., Fishell, E., Jong, R.A., Yaffe, M.J.: Symmetry of projection in the quantitative analysis of mammographic images. *European Journal of Cancer Prevention* 5, 319–327 (1996)
15. Boyd, N.F., Rommens, J.M., Vogt, K., Lee, V., Hopper, J.L., Yaffe, M.J., Pateron, A.D.: Mammographic breast density as an intermediate phenotype for breast cancer. *The Lancet Oncology* 5, 798–808 (2005)
16. Byng, J.W., Boyd, N.F., Fishell, E., Jong, R.A., Yaffe, M.J.: The quantitative analysis of mammographic densities. *Physics in Medicine and Biology* 39, 1629–1638 (1994)
17. Jain, A.K., Duin, R.P.W., Mao, J.: Statistical pattern recognition: A review. *IEEE Tr. on PAMI* 22(1), 4–37 (2000)
18. Raundahl, J., Loog, M., Nielsen, M.: Mammographic density measured as changes in tissue structure caused by hrt. *SPIE Medical Imaging* (2006)
19. Koenderink, J.J.: The structure of images. *Biological cybernetics* 50(5), 363–370 (1984)