

Note: This is an electronic reprint of the paper published in Academic Radiology as:

C.Castella, K. Kinkel, M. P. Eckstein, P.-E. Sottas, F. R. Verdun, F. O. Bochud, "Semiautomatic Mammographic Parenchymal Patterns Classification Using Multiple Statistical Features," Academic Radiology 14 (12), 1486-1499 (2007)

<http://www.academicradiology.org/article/PIIS1076633207004059/abstract>

DOI: <http://dx.doi.org/10.1016/j.acra.2007.07.014>

Semiautomatic Mammographic Parenchymal Patterns Classification Using Multiple Statistical Features

Abstract:

Rationale and Objectives. Our project was to investigate a complete methodology for the semi-automatic assessment of digital mammograms according to their density, an indicator known to be correlated to breast cancer risk. The BI-RADS four-grade density scale is usually employed by radiologists for reporting breast density, but it allows for a certain degree of subjective input, and an objective qualification of density has therefore often been reported hard to assess. The goal of this study was to design an objective technique for determining breast BI-RADS density.

Materials and Methods. The proposed semi-automatic method makes use of complementary pattern recognition techniques to describe manually selected regions of interest (ROIs) in the breast with 36 statistical features. Three different classifiers based on a linear discriminant analysis or Bayesian theories were designed and tested on a database consisting of 1408 ROIs from 88 patients, using a leave-one-ROI-out technique. Classifications in optimal feature

subspaces with lower dimensionality and reduction to a two-class problem were studied as well.

Results. Comparison with a reference established by the classifications of three radiologists shows excellent performance of the classifiers, even though extremely dense breasts continue to remain more difficult to classify accurately. For the two best classifiers, the exact agreement percentages are 76% and above, and weighted kappa values are 0.78 and 0.83. Furthermore, classification in lower dimensional spaces and two-class problems give excellent results.

Conclusion. The proposed semi-automatic classifiers method provides an objective and reproducible method for characterizing breast density, especially for the two-class case. It represents a simple and valuable tool that could be used in screening programs, training, education, or for optimizing image processing in diagnostic tasks.

Keywords: Image analysis; Pattern recognition; Feature extraction; Mammography

1. Introduction

While the etiology of breast cancer remains unclear, many studies have demonstrated a correlation between cancer risk and factors such as age, breast-feeding and pregnancy history, family history of breast cancer, hormonal treatments, genetic factors, and breast density (1-7). Breast density as a factor of risk was first investigated by Wolfe (8), who defined a four-grade density scale on the basis of the patterns and textures observed on mammograms. Later, the BI-RADS (Breast Imaging Reporting Data System) density scale was developed by the American College of Radiology to standardize mammography reporting terminology and assessment and recommendation categories (9;10). The BI-RADS density classification was created to inform referring physicians about the decline in sensitivity of mammography with increasing breast density. BI-RADS defines breast density 1 as almost entirely fatty, density 2 as scattered fibroglandular tissue, density 3 as heterogeneously dense tissue and density 4 as extremely dense tissues. It was not intended to serve as a method of measuring breast density percentage, although as per Wolfe's scale (11), correlations with this more objective factor do exist (12). In clinical American and European conditions, the breast density of a given patient is typically evaluated and reported by a radiologist using BI-RADS on the basis of the simultaneous display of two mammograms per breast.

However, one of the difficulties for correctly assessing breast density is that the BI-RADS density scale definitions are rather subjective. A certain interpretational freedom prevents perfect inter- and even intraobserver reproducibility (13;14). On the other hand, numerous pattern recognition and classification techniques have been developed and can be directly applied to this task (15). Which is why different statistical approaches have been explored in the last few years in order to develop an objective classifier of mammograms according to Wolfe or the BI-RADS scale. These techniques have made use of various pattern recognition parameters to statistically describe the whole breast or part of it: fractal dimension

(16-18), gray level histogram properties (19;20), moments (17;18;21), gray level variations matrices (17;20), or maximum response filters (22). These descriptions have been combined with several general classification algorithms: Bayesian classification (16;17), linear discriminant analysis -LDA-(20), nearest neighbour rules (21), neural networks, and textons (22).

The goal of this study was to develop a semi-automatic method for assessing the BI-RADS density category using features extracted on mammograms. For this purpose, we combined a large number of statistical features computed from manually selected regions of interest (ROIs) with linear discriminant analysis and Bayesian predictors. Special care was applied in order to assess the robustness of the three distinct classifiers we developed, and the validation of their individual performance. In contrast to most previous studies, we worked on multiple regions of interest (ROIs) per mammogram. Homogeneity in both size and emplacement was retained in order to facilitate the inter-patient comparisons of the statistical features without bias due to different breast sizes and shapes.

Each classifier was trained and tested using the leave-one-out technique to classify a set of 1408 ROIs extracted from 88 patients, on the basis of all computed features. Additionally, we averaged the individual ROIs results over multiple ROIs from the same breast and/or patient. Finally, optimal subsets of features were computed and the classifiers ran the same processes. The results were then compared to a reference classification established upon a consensus of three radiologists through weighted Kappa statistics.

The developed semi-automatic classifiers may have valuable applications in screening exam procedures, to help radiologists objectively determine breast density in a reproducible way. Patients with higher density breast tissue may thus receive special attention and specific image display optimization, since pathologies tend to be hidden by dense backgrounds. The field of potential usefulness of such classifiers extends to training and education as well.

2. Material and Methods

2.1. Mammograms database

The image database consisted of a set of 352 digital mammograms collected at the Clinique des Grangettes, Geneva, Switzerland, from patients who underwent screening exams. For each of the 88 patients, one cranio-caudal (CC) and one medio-lateral oblique (MLO) view mammogram per breast was considered. All mammograms were obtained using automatic exposure control (27-32 kV voltage) on a GE Senograph 2000D full-field digital detector (23-25). This means that not only the tube loading, but also the anode/filter combination and tube potential were selected automatically in a process involving a pre-exposure, depending on the thickness and density of the compressed breast, in order to control the dose delivered in the central breast region (26). Mammograms were outputted as 12 bits processed images, with 0.1 by 0.1 mm pixel size. All mammograms showing any sign of abnormal mammographic features such as masses, architectural distortion or clusters of microcalcification were excluded from this study.

2.2. Selection of regions of interest

The first step consisted of the manual choice of four ROIs per mammogram. The ROIs were 256 by 256 pixel square regions chosen in the central breast region, about half way between the nipple and the chest wall. One example case is shown on figure 1. The location choices were made under the control of the radiologists involved in the study and allowed us to obtain four non-overlapping ROIs per mammogram, while covering most of the breast density. This location also ensured that we performed our analysis using only breast tissue, without bias introduced by the pectoral muscle or imaging artefacts.

2.3. Statistical description

All ROIs were then characterized by the statistical quantities defined below. Unlike a global analysis of the whole breast projection, the square and uniform shape of all ROIs greatly simplifies the computation and inter-patient comparison of these features.

In order to capture as much information as possible, we extracted 18 different and complementary statistical quantities from each ROI. Due to the diversity of definitions found in the literature for a given quantity, all expressions used in this work are presented explicitly in the Appendix. They involve quantities derived from the gray level histogram like the standard deviation, skewness and kurtosis, but also balance (15;27). Gray level co-occurrence matrices (GLCM) provided quantities like energy, entropy, cmax, contrast, and homogeneity (28-30). From the primitives matrix (PM), we derived the short primitive emphasis (spe), the long primitive emphasis (lpe) as well as gray level uniformity (glu) and primitive length uniformity (plu) (28). The fractal dimension was calculated by a box-counting method (16;17;31). Finally, the neighbourhood gray-tone difference matrix (NGTDM) provided the coarseness, contrast, complexity and strength (32).

Features derived from the gray level histogram characterize the distribution of gray levels in a comprehensive way, in particular its shape and its symmetry. Balance is closely related to skewness and describes the asymmetry of the gray level histogram.

Gray level co-occurrence matrices are a powerful tool for obtaining information about the spatial relationships of gray levels in structural patterns. The ROIs were linearly re-scaled from 12 to 4 bits (16 gray levels), reducing the computing time by a factor of 65,536 and ensuring that the GLCM elements were essentially non-zero. Following, for each ROI, 20 co-occurrence matrices were computed, using directions of 0° , 45° , 90° and 135° and distances of 1, 3, 5, 7 and 9 pixels. These directions correspond to the four natural directions for a square image, and the corresponding distances describe structures from the mm to the cm

range, which are typical for the breast texture. Finally, five scalar features (energy, entropy, maximum, contrast and homogeneity) were averaged on these 20 matrices.

Primitives matrices or acquisition length parameters characterize the shape and the size of the textural patterns in an image. GLCM features are four scalars extracted from a matrix \mathbf{B} , where each element $\mathbf{B}(a,r)$ is the number of primitives of length r and gray level a , a primitive being a contiguous set of pixels with the same value. In our case, \mathbf{B} was computed from the re-scaled ROI as a 16 by 256 matrix.

Fractal dimension was calculated using the method described in detail by Caldwell (16) and Byng (17). The pixel value was seen as z-coordinate (x and y being its position in the ROI) and ruler sizes ϵ of 1 to 10 pixels were used to plot the log of the exposed surface $A(\epsilon)$ versus $\log(\epsilon)$. From this plot, the fractal dimension was computed using Eq. (26) given in the Appendix. This feature indicates the degree of complexity in the textural patterns, a low fractal dimension denoting a rather simple and homogeneous structure.

Finally, we used the textural features described by Amadasun and King (32) to get four additional statistical parameters from the NGTDM. These features provide a mathematical description of the texture and are supposed to characterize texture properties like coarseness or complexity in the same way as human observers would do. ROIs were re-scaled to 8 bits for the same reasons as for the GLCM and primitives matrices.

The statistical characterization was also performed at another scale on the same ROIs. For this, all ROIs were averaged on square blocks of 8 by 8 pixels (thus leading to 32 by 32 pixels images). All the 18 above-mentioned quantities were then computed again on these images and this provided a description of the texture at another scale, one order of magnitude higher than the first one. This step was inspired by the fact that the structures visible on mammograms are typically in the sub-mm to cm range. The total number of statistical features was thus 36, corresponding by definition to the dimension N of the classification

process. Table 1 summarizes the whole set of 18 statistical features that were computed for each of the two scales, making a total of 36 features.

2.4. Definition of gold standard from radiologists' ratings

In order to get a reliable gold standard, we asked three experienced radiologists (each of them having more than 10 years experience in radiology) to separately classify the 88 left / right pairs of CC views and the 88 pairs of MLO views mammograms, presented in random order on a laptop screen. The screen resolution was 3.6 pixels per millimetre, and brightness and contrast were adjusted before the reading session. The radiologists performed the classification individually, following the BI-RADS density scale definitions. Gold standard class was then defined for each of the 176 pairs of mammograms from the three radiologists' classifications (see Section 4).

2.5. Classification algorithms

The general purpose of pattern recognition is to determine to which category or class a given sample belongs (33). In this study, the samples are not directly the regions of interest: each ROI is characterized by an N-dimensional vector containing its computed statistical features ($N = 36$), and this observation vector serves as the input to a decision rule by which one of the given classes is attributed to the corresponding ROI. For the evaluation of the performance of the decision rule, the obtained classification is usually compared to a gold standard (also known as ground truth), which is assumed to represent the perfect classification of the samples.

All supervised classification algorithms require a set of training samples in order to establish the decision rule and a testing set to apply it. We used the leave-one-out method to avoid any bias introduced by testing on training samples. In this method, the tested ROI is always excluded from the learning process, while all other remaining ROIs are used to form

the training set. Since the ROIs were strictly non-overlapping, the 15 other ROIs selected from the same patient as the tested ROI were not excluded from the training set. This limitation allowed us to keep the number of training samples larger than N in all cases, which was a necessary condition for the computation of the features vectors covariance matrices.

We used three types of classification algorithms, namely a Bayesian classifier based on the measure of Mahalanobis distance, a Naïve Bayesian classifier and Linear Discriminant Analysis (LDA). For all methods, the samples were the N -dimensional vectors characterizing the ROIs and the four density classes were used for both training and classification phases. Concretely, each ROI (represented by its projection onto the 36-dimensional features space) was successively considered as the test ROI. The decision rules for each classifier were computed from the training set consisting of the remaining 1407 ROIs, and a density class C_R attributed to the test ROI. The procedure was repeated until a class had been given to each ROI.

2.5.1 Classical Bayesian classifier based on Mahalanobis distance

For the Bayesian classifier, 50 ROIs per density class were chosen randomly from the actual training set and thus formed four subsets $\{S_k\}_{1 \leq k \leq 4}$, each one containing 50 samples according to the gold standard (34). Assuming that the distribution of samples in each class could be approximated by an N -dimensional normal distribution, the probability of observing a given sample \mathbf{v} in the class k is given by:

$$\psi_k(\mathbf{v}) = \frac{1}{\sqrt{(2\pi)^N \det \mathbf{K}_k}} \cdot \exp\left[-\frac{1}{2}(\mathbf{v} - \boldsymbol{\mu}_k)^T \mathbf{K}_k^{-1}(\mathbf{v} - \boldsymbol{\mu}_k)\right], \quad (1)$$

where $\boldsymbol{\mu}_k$ represents the mean vector of class k and \mathbf{K}_k is the covariance matrix of vectors in class k :

$$\boldsymbol{\mu}_k = \frac{1}{n_k} \sum_{\mathbf{v}_i \in S_k} \mathbf{v}_i \quad (2)$$

$$\mathbf{K}_k = \frac{1}{n_k - 1} \sum_{\mathbf{v}_i \in S_k} (\mathbf{v}_i - \boldsymbol{\mu}_k)^T (\mathbf{v}_i - \boldsymbol{\mu}_k) \quad (3)$$

The product $(\mathbf{v} - \boldsymbol{\mu}_k)^T \mathbf{K}_k^{-1} (\mathbf{v} - \boldsymbol{\mu}_k)$ in eq. (1) is known as the square of Mahalanobis distance and is a normalized measure of the distance between the sample vector \mathbf{v} and the class center $\boldsymbol{\mu}_k$. \mathbf{K}_k and $\boldsymbol{\mu}_k$ were estimated from the sets $\{\mathbf{S}_k\}_{1 \leq k \leq 4}$ of 50 samples randomly chosen in the training set, to reduce computational cost and avoid unwanted rounding effect.

Under these assumptions, a Bayesian classifier could be defined. For a given sample \mathbf{v} , the output of the classifier was a four-dimensional vector containing the four *a posteriori* probabilities $p(k|\mathbf{v})_{1 \leq k \leq 4}$ for \mathbf{v} to belong to class k as:

$$p(k|\mathbf{v}) = \frac{p(\mathbf{v}|k)p(k)}{p(\mathbf{v})} = \frac{\psi_k(\mathbf{v})p_a(k)}{\sum_k \psi_k(\mathbf{v})p_a(k)} \quad (4)$$

The attributed class was derived from the *a posteriori* probability vector components $p(k|\mathbf{v})$ as:

$$c_R = \sum_{k=1}^4 k \cdot p(k|\mathbf{v}), \quad (5)$$

c_R being rounded to the nearest integer value to obtain the class attributed to the tested sample vector \mathbf{v} .

In eq. (4), the *a priori* probability set $\{p_a(k)\}_{1 \leq k \leq 4}$ was estimated as:

$$p_a(k) = \frac{1}{4}, \quad (6)$$

which represents the most conservative *a priori* assumption.

2.5.2 Naïve Bayesian classifier

For the second classifier, we implemented Naïve Bayesian classification, which has been proven very powerful (35), even when the assumption of feature independence given the class, which is a sufficient condition for this method to be optimal, is violated (36). The proposed normalization forced the features to be independent and also greatly simplified the computation of $p(k|\mathbf{v})$, since eq. (1) could be rewritten as:

$$\psi_k(\mathbf{v}_{n,k}) = \frac{1}{\sqrt{(2\pi)^N}} \exp \left[-\frac{1}{2} \mathbf{v}_{n,k}^T \mathbf{v}_{n,k} \right], \quad (7)$$

where \mathbf{v} has been normalized in the same way as training samples of class k to obtain the normalized vector $\mathbf{v}_{n,k}$. The four *a posteriori* probabilities $p(k|\mathbf{v})$ were then computed with eq. (4), and the attributed class with eq. (5).

We thus modified the Bayesian classifier procedure so that all feature distributions were within-class normalized. In order to force a distribution to be normal, its cumulative histogram was compared to the integral of the Gaussian density function: the normalized value $p_n^{j,k}$ of a given parameter $p^{j,k}$ is the solution of the equation:

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{p_n^{j,k}} e^{-t^2/2} dt = \frac{p^{j,k}}{p_{max}^{j,k}}, \quad (8)$$

where $p_{max}^{j,k}$ is the highest value in the original distribution of feature j in class k .

2.5.3 Linear Discriminant Analysis

Linear Discriminant Analysis implemented in Matlab Statistics Toolbox (37) is essentially similar to the first described algorithm, except that in Eq. (3) only one pooled covariance matrix \mathbf{K}_o is computed instead of the four \mathbf{K}_k (homoscedasticity hypothesis), forcing the borders in the features space to be hyperplanes instead of quadrics. A multivariate normal density is then fitted to each class:

$$\psi_k(\mathbf{v}) = \frac{1}{\sqrt{(2\pi)^N \det \mathbf{K}_o}} \cdot \exp \left[-\frac{1}{2} (\mathbf{v} - \boldsymbol{\mu}_k)^T \mathbf{K}_o^{-1} (\mathbf{v} - \boldsymbol{\mu}_k) \right] \quad (9)$$

Following, the decision rule used to attribute a class to a sample is in this case a simple linear combination of the features (38). The LDA classifier returns the class C_R corresponding to its position in the features space for each tested sample. This means that the *a posteriori* vector had only one non-zero component. As opposed to the classical Bayesian classifier described in Section 2.5.1, this variant made use of all ROIs present in the training set, without having to define one subset S_k per class.

2.5.4 Averaging the individual ROIs classifications

These three classifiers were used to individually classify all 1408 ROIs. However, since the BI-RADS density scale is based on an overall appreciation of the breast and since an overall

dense breast may contain one or several ROIs that are essentially fatty, individual ROI classification may lead to results that differ from the radiologist' evaluations. Therefore, we also introduced two kinds of averaging to avoid decisions that were too localized. First, *a posteriori* probability vectors $[p(k|v)_{1 \leq k \leq 4}]$ were averaged for each mammogram over the four corresponding ROIs and eq. (5) was used again to attribute a general class to each mammogram, instead of one per ROI. Secondly, we studied the effect of averaging on the 8 ROIs (four per mammogram) that had been defined for each left / right pair of CC or MLO views. This corresponds to the situation nearest to that of the three involved radiologists, who established the gold standard based on the display of a left / right pair of mammograms.

2.5.5 Reduction of the features space size and number of classes

In order to reduce the original dimensionality of the features vector ($N=36$) to a given $N' < N$ and to determine for that given N' which parameters would lead to the best classification performance, we used standard features extraction techniques based on the maximization of the between-class scatter to the within-class scatter (Fisher Linear Discriminant) (39-41). Concretely, the Fisher Linear Discriminant gives a measure of the separability of the four density classes when only N' features amongst the original N ones are considered for the classification. This process was conducted for $N'=2$ and 5, and the separability measure was computed for every combination of N' parameters (brute force testing). Once the best combination had been identified, all previously described algorithms were applied to the feature vectors orthogonally projected on the obtained subspaces, meaning that the classifiers only used the N' best features for defining their classification rules.

We also examined the case of grouping BI-RADS 1 and 2 in the same density class, and BI-RADS 3 and 4 in another. We compared the performance obtained with this grouping being done before the training process, or after the classification (thus respectively 2-class training – 2-class classification and 4-class training – 2-class classification).

2.5.6 Evaluation of the performance

We used Kappa statistics with quadratic weights to evaluate the performance of the classification algorithms (42-46). This parameter represents the degree of chance-corrected agreement between two classifications (classification algorithm versus gold standard or radiologist versus radiologist) as:

$$\kappa = \frac{p_o - p_e}{1 - p_e}, \quad (10)$$

where p_o is the observed agreement proportion and p_e the agreement expected by chance alone. Both are calculated from the confusion matrix and the quadratic weights matrix, and the values of κ stand between -1 and 1 (the minimum value actually depends on p_e but is always between -1 and 0). Benchmarks by Landis and Koch (47) (adjusted by Fleiss *et al* (42) for taking the weighting process into account) are commonly used: Kappa values below 0.4 reflect poor agreement, between 0.4 and 0.6 moderate agreement, while it is substantial between 0.6 and 0.75 and excellent above 0.75. Weighted Kappa is particularly well adapted to multi-class tasks and when the classes are rather subjectively defined, which is the case for the BI-RADS density scale. The weighting process indeed differentiates between serious (more than one BI-RADS class difference) and slight disagreement (immediate neighbour class choice), and has been chosen as an evaluation parameter in numerous previous works on mammogram classification (16;17;20) Although much more sensitive to differences in class prevalence, the exact agreement proportion was also computed to be able to compare the performance with results from other studies (16;21;22)

3. Results

The reference classifications by the three radiologists involved in this study are summarized in table 2 and figure 2. The exact agreement among the three classifications was 55%, while for the remaining 45% two of the radiologists chose a given BI-RADS density class and the last one chose an immediate neighbouring class. When compared with each other, the three

radiologists involved in our study obtained 67% to 79% exact agreement. The values of Kappa and exact agreement percentage, for each radiologist versus gold standard, are summarized in table 2. Figure 2 presents the number of cases per radiologists' consensus level. The latter is defined as the number of radiologists having chosen the same BI-RADS category.

Typical time periods to train and test the classifiers were 90 minutes for Naïve Bayesian, 5 minutes for the Mahalanobis-Bayesian and 1 minute for LDA classifier, on a Pentium 4 (3GHz processor, 512 MB RAM). In the 36-dimensional feature space, Naïve Bayesian classification led to a Kappa value of 0.68 ± 0.07 and a percentage agreement with respect to the gold standard of 60%. This classifier was outperformed by the two others, since we obtained Kappa values of 0.78 ± 0.07 for Mahalanobis-Bayesian and 0.83 ± 0.08 for LDA. As one can expect from the overlap of standard errors, paired t-tests showed that none of these differences were significant at the 5% confidence level. The exact agreement proportions between these classifiers and the gold standard were respectively 76% and 83%. The confusion matrices given in table 3 for the two best classifiers show that all but one mammogram pair were classified in the correct class or in one of its immediate neighbours. Moreover, this result was also valid when comparing breast density assessment of individual breasts before averaging the left / right pairs. The effect of the averaging process (individual ROI classification, averaging over the four ROIs defined for each mammogram, and averaging over the 8 ROIs defined on a left / right pair of mammograms) is presented in table 4.

The dimensionality reduction to $N'=2$ and 5 has as expected an effect on classification performance. As shown on figure 3, Kappa decreases when the number of features is reduced, although both methods obtain already good results with 5 parameters only. The two optimal features for differentiating the four classes were homogeneity and coarseness, and the

corresponding partition of the bi-dimensional subspace is given in figure 4. For $N'=5$, the optimal parameters were standard deviation, skewness, primitive length uniformity, fractal dimension, and coarseness, the latter parameter being computed from the block-averaged and the first four from the original ROI.

The reduction to a 2-class problem led to the same results when the grouping of BI-RADS density classes was done before or after training. Naïve Bayesian classifier obtained kappa values and percentage agreement of 0.68 ± 0.08 and 86%. Even if the difference is not significant at the 5% confidence level, it was once again outperformed by Mahalanobis Bayesian and LDA classifiers, for which the exact agreement were respectively 88% and 90% and weighted kappa were 0.74 ± 0.08 and 0.78 ± 0.08 . Thus the performances of the last two classifiers for that particular two-class problem are excellent and nearly equivalent.

Finally, we observed no difference between the results obtained for CC and MLO views: performance of the automatic classifiers remained unchanged when the training phase was performed on one type of view and the classification on the other, or when training and classification processes were restricted to one view.

4. Discussion

Since BI-RADS scale definitions allow for a certain freedom regarding interpretation, it was essential to carefully define the gold standard. The number of radiologists devoted to that task was between one and four among other published studies (16;17;20;21). The choice of three radiologists for this study was adequate, in the sense that there was no case where the three radiologists chose three different classes, or where one would have chosen a non-immediate neighbouring class respectively to the others. Thus the odd number of radiologists permitted in all cases to unequivocally define the gold standard classification, as the class selected by at least two radiologists. The different case repartitions among the four BI-RADS classes are shown on figure 2. The first radiologist tended to use the lowest categories more often than

the other two. The second observer classified the same proportion of mammograms between BI-RADS 1 and 2 categories, while reporting more than 60% in BI-RADS 3 category. The third observer barely used the extreme categories and concentrated most answers in BI-RADS 3 category as well.

The choice of presenting CC and MLO views separately to the radiologists allowed us to show that intra-observer reproducibility was excellent, even for different X-ray projections. The proportions of cases with one class difference between CC and MLO classifications were respectively 14%, 15%, and 9% for radiologist #1, #2 and #3. No difference greater than one BI-RADS density class was observed. Thus the corresponding confusion matrices (observer i CC classification vs. observer i MLO classification) led to very high weighted kappa values (0.90, 0.87, 0.87), showing that radiologists' classifications were nearly independent from the presented view. However, it was observed that the first observer attributed one class higher to MLO compared to CC for 10 out of its 12 differences, while the second had the opposite trend (one class higher for CC view for 9 out of the 13 differences), and the third observer had roughly equally distributed differences (5 cases out of 8 with one class higher for CC).

The analysis of each within-class features distributions was in total agreement with the intuitive meaning of the statistical parameters and the two-scale analysis on normal and block-averaged ROIs provided very coherent results: the same trends were observed at mm and cm scales. Texture elements in low density breasts are small, fine and well contrasted, with a high fractal dimension, while patterns in high density breasts are much coarser, due to the diffusive nature of glandular tissues.

The Naïve Bayesian classifier obtained substantial agreement, but as some of the 36 features were strongly correlated, its performance was degraded as expected (35). The results of LDA and Bayesian classification based on the measure of Mahalanobis distance, in the 36-dimensional feature space, were remarkable, with cross-validated Kappa values of 0.78 and

0.83 respectively, and exact recognition proportions of 76% and 83%. LDA's slightly better performance is probably due to the fact that the whole 1407 ROIs training set was used for establishing the classification rules, whereas the same number of samples per density class, 50, was used for the Bayesian classifier, in order to avoid over-training in the most represented classes. The confusion matrices given in table 3 show an excellent differentiation of the four classes. However, half of the BI-RADS 4 cases were misclassified in density category 3 or 2. This may indicate that the sample size was too small for this category, or that the gold standard assessment for this category was not accurate enough.

Compared to previous studies using a four-grade scale (Wolfe scale (Caldwell *et al* (16), Tahoces *et al* (20)), dense breast tissue proportion (Karssemeijer (21)), BI-RADS (Petroudi *et al* (22)), an improvement of Kappa and recognition rate was obtained: Caldwell cites indeed Kappa values between 0.58 and 0.61, Tahoces between 0.63 and 0.71, Karssemeijer a global value of 0.73 and Petroudi a recognition rate of 76%. This improvement can probably be explained by the homogeneity of the ROIs in size and localization in the breast, the absence of any background- or pectoral muscle removal algorithm, the use of digital mammograms instead of digitized films, the wide range of complementary texture analysis techniques, and the averaging processes to take into account most of the dense breast region. Comparisons with these studies should proceed carefully, however, since little information is mentioned about pre-processing of the mammograms or case-distributions. In addition, the comparison between the other scales with BI-RADS classification results is not a trivial point.

The reduction to a 2-grade scale (BI-RADS 1-2 vs BI-RADS 3-4, 88% recognition rate and higher) led to an excellent performance as well, comparable to the results given by Bovis and Singh (18).

When the algorithm had to make its decision based on two or five statistical parameters only, we found a substantial (weighted kappa > 0.7) to excellent agreement with respect to the gold standard. The Naïve Bayesian classifier proved that its performance is excellent for low-dimensional feature spaces, where the independence assumption can still be considered as valid. The performance of Mahalanobis Bayesian and LDA classifiers increased with the dimension of the optimal features subspace, with slightly better results for LDA for 5 features and above. It is interesting to note how the most optimal parameters were chosen in order to be complementary. For instance, with 5 parameters, two features related to gray level histogram (standard deviation and skewness), one from PM (plu), one from NGTDM (coarseness), and the fractal dimension were selected. This complementary nature between all texture analysis methods is one of the key points for obtaining a good classification even in a low-dimensional features space.

The improvement gained when averaging the results over four ROIs defined in the same breast and over the left / right breasts pair is clear for all classifiers. It shows that this process is the best way to take into account a significant part of both breasts, and thus avoid making a too local decision. Local classification, as shown in table 4 for individual ROIs results, is not efficient for both Bayesian classifiers, although already substantially good for LDA.

Finally, according to Karssemeijer (21), the upper limit of the performance of an automatic classifier in terms of comparison with human observers, remains an open question. It would be interesting to compare the gold standard defined in this study with other independent radiologists' classifications to have an idea of an empirical value of maximum Kappa and the exact agreement one could expect, the latter being evaluated by Karssemeijer (21) to be 80%. The exact agreement between the three radiologists involved in our study when compared with each other (67% to 79%) lies effectively in this range.

5. Conclusion

An excellent assessment of breast density according to BI-RADS was obtained with the semi-automated method presented in this study. A complete method was used combining complementary methods (moments, GLCM, PM, fractal dimension and NGTDM) to describe ROIs manually chosen on digital mammograms, with widely used classification methods (LDA, Bayesian classification) and different averaging processes in order to take into account as much comprehensive information as possible. The results showed that the agreement between the radiologists and the automatic classifiers was notably higher than most previous published values, although extremely dense breasts (BI-RADS category 4) seemed somehow more difficult to classify accurately. Using $N=36$ parameters led to high performance for the assessment of designing an automatic breast density classifier. The usefulness of mixing complementary methods was demonstrated by reducing the dimensionality of the feature space to 5 optimal parameters. The classifiers obtained excellent performances as well when tested in the two-class problem reduction. In a future phase, the validation procedure, currently limited to leave-one-ROI-out and justified by the fact that the ROIs do not overlap, could be extended to leave-one-patient-out on a larger patient database. The excellent results obtained with the most represented classes (BI-RADS categories 2 and 3) and with cross-views training and testing, suggest that the bias introduced by the leave-one-ROI-out method, if any, should not influence the overall performance of the classifiers, since in these cases training and testing on ROIs that had been taken from the same mammogram was less likely to happen.

The other key feature of the method resides in its simplicity. Apart from the fast computation of the 36 parameters, no additional algorithm is needed to remove the background, the pectoral muscle and any potential imaging artefact, since a total control over the location of the ROIs is kept by manually selecting them. A fully automated classifier with

a built-in location selection algorithm has not been investigated in this paper, but existing breast segmentation methods (12;48) could certainly be combined with the proposed classifiers to improve reproducibility and accuracy of the locations choices. The automatisation of ROI selection would help build a larger, more objective database, which is currently the main limitation of this study.

The proposed method represents a valuable tool for use in screening programs and could be inserted in a CAD device, in order to help radiologists in their density evaluation and diagnosis tasks. Intra- or inter-observer variability in density assessment could indeed be avoided through the help of an automatic or semi-automatic classifier, and optimized data processing could be applied in order to display an optimal image to the radiologists for their diagnosis. An objective tool for determining breast density may find other potential applications in follow-up management for patients, with screening frequencies depending on breast density. Finally, training and education may benefit from such classifiers, in order to lower the variability of intra- and inter-observer classifications inherent to the BI-RADS density class definitions.

Acknowledgments

The authors are grateful to Elsabe Scott, MD, and Nigel Howarth, MD, for the classification of the mammograms, and to Samuel De Laere for collecting the mammograms and creating the database. This work was supported by Swiss National Science Foundation FN 3252BO-104273.

Appendix: Definition of the statistical parameters

A. Parameters computed from the gray level histogram

The first four moments and balance parameter are computed from the individual pixel values x_i as follows:

$$\text{mean} \equiv \bar{x} = \frac{1}{N} \sum_i x_i \quad (11)$$

$$\text{standard dev.} \equiv \sigma = \frac{1}{\sqrt{N-1}} \left(\sum_i (x_i - \bar{x})^2 \right)^{1/2} \quad (12)$$

$$\text{skewness} = \frac{1}{N\sigma^3} \sum_i (x_i - \bar{x})^3 \quad (13)$$

$$\text{kurtosis} = \frac{1}{N\sigma^4} \sum_i (x_i - \bar{x})^4 - 3 \quad (14)$$

$$\text{balance} = \frac{x_{70} - \bar{x}}{\bar{x} - x_{30}}, \quad (15)$$

where the summations are performed over the N pixels of the ROI, and x_p is the gray level yielding to p -th percentile of the gray level distribution (15)

B. Gray level co-occurrence matrices (GLCM)

The GLCM are computed as follows: first, the ROI is linearly re-scaled to 16 gray levels only. Then for a given direction d and a given distance r , each element $[i,j]$ of the co-occurrence matrix $(\mathbf{C})_{i,j}^{dr}$ is given by the number of times that a couple of pixels separated by a distance r along a direction d have the values i and j respectively. Each co-occurrence matrix is then normalized by the sum of its elements. The directions chosen for the GLCM are $[1,0]$, $[1,1]$, $[0,1]$ and $[-1,1]$, corresponding to angles of 0° , 45° , 90° and 135° respectively. The distances are 1, 3, 5, 7 and 9 pixels for each direction, which yields to a set of 20 GLCM. Scalar parameters are then extracted from each matrix as follows:

$$\text{energy}(\mathbf{C}) = \sum_{i,j} \mathbf{C}_{i,j}^2 \quad (16)$$

$$\text{entropy}(\mathbf{C}) = - \sum_{i,j} \mathbf{C}_{i,j} \log \mathbf{C}_{i,j} \quad (17)$$

$$\text{cmax}(\mathbf{C}) = \max_{i,j} \mathbf{C}_{i,j} \quad (18)$$

$$\text{contrast}(\mathbf{C}) = \sum_{i,j} |i-j|^2 \mathbf{C}_{i,j} \quad (19)$$

$$\text{homogeneity}(\mathbf{C}) = \sum_{i,j} \frac{\mathbf{C}_{i,j}}{1+|i-j|} \quad (20)$$

C. Primitives matrix (PM)

Each element $[a,r]$ of the primitives matrix $\mathbf{B}_{a,r}$ is the number of primitives of gray-level a and length r , a primitive being a contiguous set of pixels having the same value. As for GLCM, each ROI is re-scaled to 4 bits before its primitives matrix is computed. Note that its dimensions are $(2^4-1, r_{max})$, since $0 \leq a \leq 2^4-1$ and $1 \leq r \leq r_{max}$, where r_{max} is the dimension of the ROI, corresponding to the maximal primitive length one could find in such an image. From this primitives matrix, four parameters are then extracted for each ROI: short primitive emphasis (spe), long primitive emphasis (lpe), gray level uniformity (glu) and primitive length uniformity (plu), defined by:

$$\text{spe} = \frac{1}{\mathbf{B}_{tot}} \sum_a \sum_r \frac{\mathbf{B}_{a,r}}{r^2} \quad (21)$$

$$\text{lpe} = \frac{1}{\mathbf{B}_{tot}} \sum_a \sum_r \mathbf{B}_{a,r} r^2 \quad (22)$$

$$\text{glu} = \frac{1}{\mathbf{B}_{tot}} \sum_a \left(\sum_r \mathbf{B}_{a,r} \right)^2 \quad (23)$$

$$\text{plu} = \frac{1}{\mathbf{B}_{tot}} \sum_r \left(\sum_a \mathbf{B}_{a,r} \right)^2, \quad (24)$$

where \mathbf{B}_{tot} is the sum of the elements of the primitives matrix \mathbf{B} : $\mathbf{B}_{tot} = \sum_a \sum_r \mathbf{B}_{a,r}$.

Note that \mathbf{B} could be defined for several directions, but we limited our investigations to one (34), corresponding to a scan of the image along direction $[1,0]$.

D. Fractal dimension

The fractal dimension of a 2-dimensional (2-D) image can be computed by a box-counting method as an extension to the 1-dimensional (1-D) case. Mandelbrot (49) first described the

1-D problem of measuring a coastline on a map, with a ruler of a particular length ϵ . The smaller the ruler, the larger the measured distance, since more and more details can be taken into account for the analysis. Mandelbrot gave the empirical relationship between the ruler size ϵ , and the measured length L , as:

$$L(\epsilon) = \lambda \epsilon^{1-D} \quad (25)$$

In Eq. (25), λ is a scaling constant, and D is called the fractal dimension of the curve.

The generalization to a 2-D image can be done as follows (16;17;31). First, the image to be analyzed is converted to a pseudo-3D surface, with the first two coordinates representing the spatial position of each pixel, the third one being the gray level. The total area A of the 3-D surface is then computed. For various values of the ruler size ϵ , the pixel values are then averaged over blocks of size ϵ by ϵ , and the area $A(\epsilon)$ is computed. For the 2-D case, Eq. (25) becomes:

$$A(\epsilon) = \lambda \epsilon^{2-D} \quad (26)$$

According to this equation, D can be estimated from a plot of $\log\{A(\epsilon)\}$ versus $\log\{\epsilon\}$.

E. Neighbourhood gray-tone difference matrix (NGTDM)

NGTDM is a column matrix first defined by Amadasun and King (32) as follows: let $x_{k,l}$ be the gray level value of the pixel located at (k,l) on a two-dimensional image. The average neighbouring value is given by:

$$\bar{A}_{x_{k,l}} = \frac{1}{W-1} \left[\sum_{m=-d}^d \sum_{n=-d}^d x_{k+m,l+n} \right], (m,n) \neq (0,0), \quad (27)$$

where $d = 3$ is the neighbouring size and $W = (2d+1)^2$. Denoting $\{X_i\}$ the set of all pixels with value i in the ROI, the i -th entry of the NGTDM is given by:

$$s(i) = \sum_{x \in X_i} |i - \bar{A}_x| \quad (28)$$

Scalar parameters extracted from the NGTDM are:

$$\text{coarseness} = \left[\varepsilon + \sum_{i=0}^{i_{\max}} p_i s(i) \right]^{-1} \quad (29)$$

$$\text{contrast}' = \left[\frac{1}{N_g (N_g - 1)} \sum_{i=0}^{i_{\max}} \sum_{j=0}^{j_{\max}} p_i p_j (i - j)^2 \right] \cdot \left[\frac{1}{n^2} \sum_{i=0}^{i_{\max}} s(i) \right] \quad (30)$$

$$\text{complexity} = \sum_{i=0}^{i_{\max}} \sum_{j=0}^{j_{\max}} \frac{|i - j| [p_i s(i) + p_j s(j)]}{n^2 (p_i + p_j)}, p_i > 0, p_j > 0 \quad (31)$$

$$\text{strength} = \frac{\sum_{i=0}^{i_{\max}} \sum_{j=0}^{j_{\max}} (p_i + p_j) (i - j)^2}{\varepsilon + \sum_{i=0}^{i_{\max}} s(i)}, p_i > 0, p_j > 0, \quad (32)$$

where $p_i = |X_i| / \sum_{i=0}^{i_{\max}} |X_i|$ is the probability of occurrence of gray level i in the ROI, i_{\max} the highest gray level and N_g the number of different gray levels effectively present in the ROI and ε a small number (10^{-12} in our case) to prevent coarseness and strength becoming infinite. The feature representing the contrast given by Eq. (30) is called here *contrast'*, to make a distinction with the contrast derived from the primitives matrices (see Eq. (19)).

Reference List

1. Fitzgibbons PL, Page DL, Weaver D, Thor AD, Allred DC, Clark GM, et al. Prognostic factors in breast cancer. College of American Pathologists Consensus Statement 1999. *Arch Pathol Lab Med* 2000;124:966-78.
2. Ziv E, Smith-Bindman R, Kerlikowske K. Mammographic Breast Density and Family History of Breast Cancer. *Journal of the National Cancer Institute* 2003;95:556-8.
3. Colditz GA, Hankison SE, Hunter DJ, Willett WC, Manson JE, Stampfer MJ, et al. The use of estrogens and progestins and the risk of breast cancer in postmenopausal women. *N Engl J Med* 1995;332:1589-93.
4. Kelsey JL, Gammon MD, John EM. Reproductive factors and breast cancer. *Epidemiol Rev* 1993;15:36-47.
5. Boyd NF, Byng JW, Long RA, Little EK, Miller AB, Lockwood GA, et al. Quantitative classification of mammographic densities and breast cancer risk: results from the Canadian National Breast Screening study. *J Nat Cancer Inst* 1995;87:670-5.
6. van Gils CH, Hendriks JH, Holland R, Karssemeijer N, Otten JD, Straatman H, et al. Changes in mammographic breast density and concomitant changes in breast cancer risk. *Eur J Cancer Prev* 1999;8:509-15.

7. Heine JJ, Malhotra P. Mammographic Tissue, Breast Cancer Risk, Serial Image Analysis, and Digital Mammography: Part 1. Tissue and Related Risk Factors. *Acad Radiol* 2001;9:298-316.
8. Wolfe JN. Breast patterns as an index of risk for developing breast cancer. *Am J Roentgenol* 1976;126:1130-7.
9. American College of Radiology (ACR) 2004. ACR Practice Guideline for the performance of screening mammography. *Practice Guidelines & Technical Standards*. 2004.
10. Harvey JA, Bovbjerg VE. Quantitative Assessment of Mammographic Breast Density: Relationship with Breast Cancer Risk. *Radiology* 2004;230:29-41.
11. Brisson J, Diorio C, Mâsse B. Wolfe's Parenchymal Pattern and Percentage of the Breast with Mammographic Densities: Redundant or Complementary Classifications? *Cancer Epidemiology, Biomarkers & Prevention* 2003;12:728-32.
12. Perconti P, Loew M. Analysis of parenchymal patterns using conspicuous spatial frequency features in mammograms applied to the BI-RADS density rating scheme. *SPIE Medical Imaging: Image Processing* 2006.
13. Kerlikowske K, Grady D, Barclay J, Frankel SD, Ominsky SH, Sickles E, et al. Variability and Accuracy in Mammographic Interpretation Using the American College of Radiology Breast Imaging Reporting and Data System. *J Nat Cancer Inst* 1998;90:1801-9.

14. Berg WA, Campassi C, Langenberg P, Sexton MJ. Breast Imaging Reporting and Data System: Inter- and Intraobserver Variability in Feature Analysis and Final Assessment. *Am J Roentgenol* 2000;174:1769-77.
15. Huo Z, Giger ML, Wolverton DE, Zhong W, Cumming S, Olopade OI. Computerized analysis of mammographic parenchymal patterns for breast cancer assessment: Feature selection. *Med Phys* 2000;27:4-12.
16. Caldwell CB, Stappleton SJ, Holdsworth DW, Jong RA, Weiser WJ, Cooke G, et al. Characterisation of mammographic parenchymal pattern by fractal dimension. *Phys Med Biol* 1990;35:235-47.
17. Byng JW, Boyd NF, Fishel E, Jong RA, Yaffe MJ. Automated analysis of mammographic densities. *Phys Med Biol* 1996;41:909-23.
18. Bovis K, Singh S. Classification of mammographic breast density using a combined classifier paradigm. *Proceedings of the 4th International Workshop on Digital Mammography* 2002 p. 177-80.
19. Zhou C, Chan HP, Petrick N, Helvie A, Goodsit MM, Sahiner B, et al. Computerized image analysis; Estimation of breast density on mammograms. *Med Phys* 2001;28:1056-69.
20. Tahoces PG, Correa J, Souto M, Gómez L, Vidal JJ. Computer-aided diagnosis: the classification of mammographic breast parenchymal patterns. *Phys Med Biol* 1995;40:103-17.

21. Karssemeijer N. Automated classification of parenchymal patterns in mammograms. *Phys Med Biol* 1998;43:365-78.
22. Petroudi S, Kadir T, Brady M. Automatic Classification of Mammographic Parenchymal Patterns: a statistical Approach. *Proc.of IEEE International Conference on Engineering in Medicine and Biology* 2003 p. 798-801.
23. Vedantham S, Karellas A, Suryanarayanan S, Albagli D, Han S, Tkaczyk EJ, et al. Full breast digital mammography with an amorphous silicon-based flat panel detector: Physical characteristics of a clinical prototype. *Med Phys* 2000;27:558-67.
24. Burgess A. On the noise variance of a digital mammography system. *Med Phys* 2004;31:1987-95.
25. Muller S. Full-field digital mammography designed as a complete system. *European Journal of Radiology* 1999;31:25-34.
26. Hemdal B, Andersson I, Grahn A, Hakansson M, Ruschin M, Thilander-Klang A, et al. Can the average glandular dose in routine digital mammography screening be reduced ? A pilot study using revised image quality criteria. *Radiation Protection Dosimetry* 2005;114:383-8.
27. Li H, Giger ML, Olopade OI, Margolis A, Lan L, Chinander MR. Computerized texture analysis of mammographic parenchymal patterns of digitized mammograms. *Acad Radiol* 2005;12:863-73.

28. Sonka M, Hlavak V, Boyle R. Image processing, Analysis and Machine Vision. 2nd ed. Pacific Grove, CA: Brooks/Cole, 1999.
29. Tuceryan M, Jain AK. Texture Analysis. In: Chen CH, Pau LF, Wang P, editors. The Handbook of Pattern Recognition and Computer Vision. River Edge, NJ: World Scientific Publishing, 1998.
30. Haralick RM, Shanmugam K, Dinstein I. Textural Features for Image Classification. IEEE Trans Syst , Man, Cybern 1973;3:610-62.
31. Lundhal T, Ohley WJ, Kunklinski WS, Williams DO, Gewirtz H, Most AS. Analysis and interpolation of angiographic images by use of fractals. Proceedings of IEEE Conference on Computers in Cardiology 1985 p. 355-8.
32. Amadasun M, King R. Textural Features Corresponding to Textural Properties. IEEE Trans Syst , Man, Cybern 1989;19:1264-74.
33. Fukunaga K. Introduction to Statistical Pattern Recognition. 2nd ed. San Diego, CA: Academic Press, 1990.
34. Chabat F, Guang-Zhong Y, Mansell DM. Obstructive Lung Diseases: Texture classification for Differentiation at CT. Radiology 2003;228:871-7.
35. Friedman N, Geiger D, Goldszmidt M. Bayesian Network Classifier. Machine Learning 1997;29:131-63.

36. Domingos P, Pazzani MJ. On the optimality of the Simple Bayesian Classifier under Zero-One Loss. *Machine Learning* 1997;29:103-30.
37. Matlab Statistics Toolbox [computer program]. Naticks: The MathWorks,Inc; 2004.
38. Matlab Statistics Toolbox [computer program]. Naticks: The MathWorks,Inc; 2004.
39. Nasri M, El Hitmy M. Algorithme génétique et Critère de la Trace pour l'Optimisation du vecteur Attribut: Application à la Classification Supervisée des Images de Textures. 15th International Conference on Vision Interface 2002.
40. Yeom S, Javidi B. Three-dimensional distortion-tolerant object recognition using integral imaging. *Optics Express* 2004;12:5795-809.
41. Barrett HH, Myers KJ. *Foundations of Image Science*. Wiley, 2004.
42. Fleiss JL, Levin B, Paik MC. *Statistical Rates and Proportions*. 3rd ed. Hoboken, NJ: Wiley, 2003.
43. Ker M. Issues in the use of Kappa. *Invest Radiol* 1991;26:78-83.
44. Kundel HL. Measurement of Observer Agreement. *Radiology* 2003;228:303-8.
45. Kraemer HC. Extension of the Kappa Coefficient. *Biometrics* 1980;36:207-16.

46. Kraemer HC, Periyakoil VS, Noda A. Kappa coefficients in medical research. In: D'Agostino RB, editor. *Tutorials in Biostatistics, vol 1: Statistical Methods in Clinical Studies*. Hoboken, NJ: Wiley, 2004.
47. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1973;33:671-9.
48. Petroudi S, Brady M. Breast Density Segmentation Using Texture. *Proceedings of the 8th International Workshop on Digital Mammography Berlin Heidelberg*: Springer; 2006 p. 609-15.
49. Mandelbrot BB. *The Fractal Geometry of Nature*. San Francisco, CA: W. H. Freeman, 1982.

List of figures

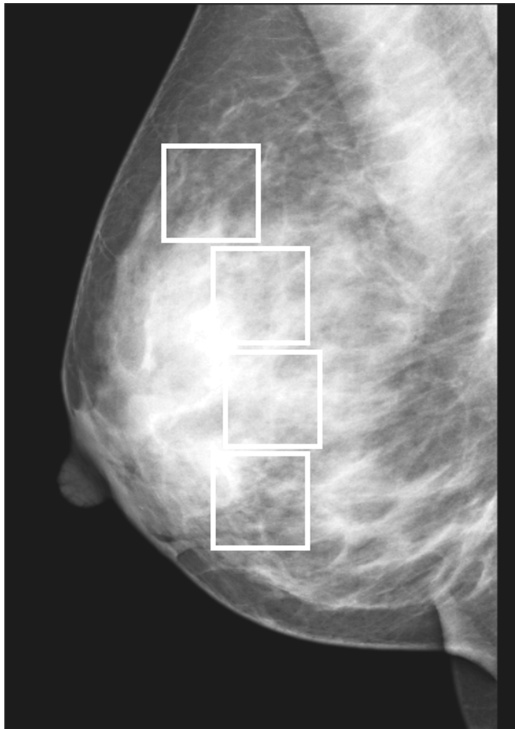


Fig. 1. Digital mammogram and corresponding manually-defined regions of interest.

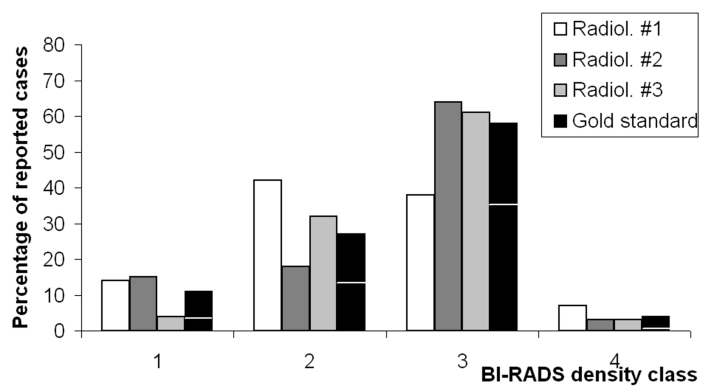


Fig. 2. Repartition of the 176 breast pairs among BI-RADS density classes. The separation line in the gold standard column indicates the proportion of cases per consensus level: 3/3 (lower part of the column) or 2/3 (upper part).

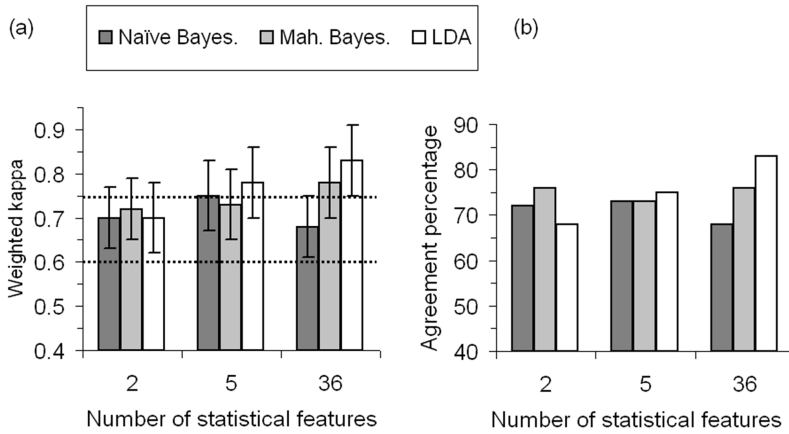


Figure 3. (a) Weighted Kappa value as a function of the features space dimensionality. Lines at 0.6 and 0.75 represent the limits for substantial and excellent agreement (b) Corresponding percentage agreement.

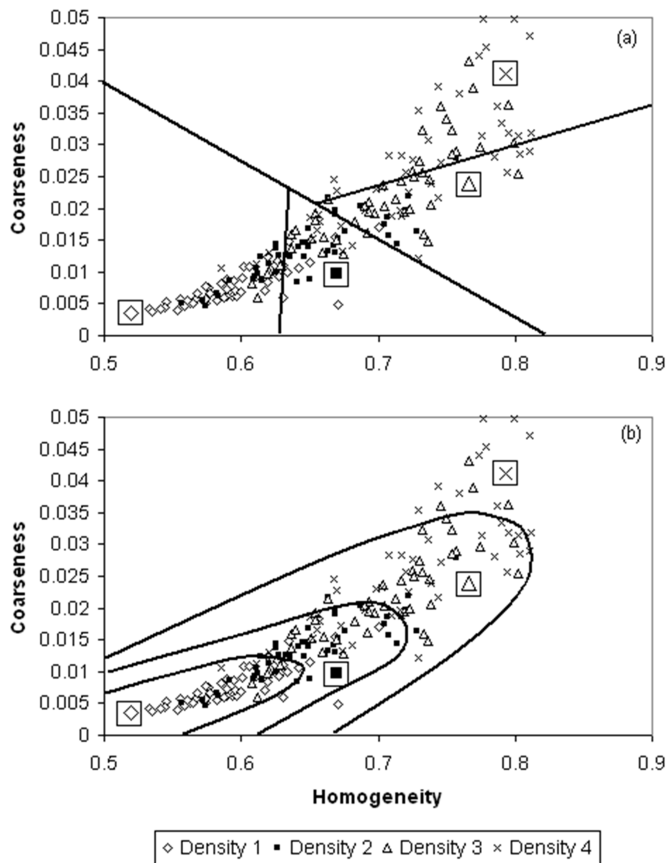


Figure 4. Partition of the optimal bi-dimensional feature subspace. (a) LDA leads to linear borders (b) For Bayesian classifier based on Mahalanobis distance, the borders are conics. For visibility reasons, only 40-50 randomly chosen ROIs per density class are shown.

List of tables

Table 1. Summary of the texture analysis methods and the corresponding features. The 18 parameters in this table were computed for two scales as described in Section 2.3, making a total of 36 features.

Analysis method	Statistical features
Gray level histogram	standard deviation
	skewness
	kurtosis
	balance
Gray level co-occurrence matrices	energy
	entropy
	cmax
	contrast
	homogeneity
Primitives matrices	short primitive emphasis
	long primitive emphasis
	gray level uniformity
	primitive length uniformity
Fractal analysis	fractal dimension
Neighbourhood gray-tone difference matrix	coarseness
	contrast'
	complexity
	strength

Table 2. Radiologist classifications compared to the gold standard classification defined in Section 4. Standard error for weighted Kappa was computed according to the formula given by Fleiss *et al* (2003).

	Radiologist #1	Radiologist #2	Radiologist #3
Kappa	0.81±0.07	0.88±0.07	0.91±0.08
Exact agreement	77%	89%	89%

Table 3. (a) Confusion matrix obtained for the Bayesian classifier based on Mahalanobis distance. Results are averaged over mammogram pairs from the same view. (b) Same for LDA classifier.

		Gold standard			
		Density 1	Density 2	Density 3	Density 4
Bayesian Classifier	Density 1	14	3	0	0
	Density 2	5	30	6	1
	Density 3	0	14	86	3
	Density 4	0	0	10	4

		Gold standard			
		Density 1	Density 2	Density 3	Density 4
LDA Classifier	Density 1	16	3	0	0
	Density 2	3	31	4	1
	Density 3	0	13	95	3
	Density 4	0	0	3	4

Table 4. Weighted kappa values obtained with the different averaging processes and classifiers.

	Individual ROI classification	Average per mammogram (4 ROIs)	Average per view type (8 ROIs)
Naïve Bayesian	0.50±0.02 (39%)	0.65±0.05 (55%)	0.68±0.07 (60%)
Mahalanobis Bayesian	0.58±0.03 (53%)	0.73±0.05 (69%)	0.78±0.07 (76%)
LDA	0.71±0.03 (70%)	0.81±0.05 (80%)	0.83±0.08 (83%)