A Novel Breast Tissue Density Classification Methodology

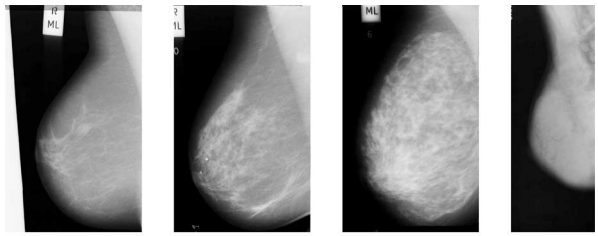
Arnau Oliver, Jordi Freixenet, Robert Martí, Josep Pont, Elsa Pérez, Erika R. E. Denton, and Reyer Zwiggelaar

***Abstract*—It has been shown that the accuracy of mammographic abnormality detection methods is strongly dependent on the breast tissue characteristics, where a dense breast drastically reduces detection sensitivity. In addition, breast tissue density is widely accepted to be an important risk indicator for the development of breast cancer. Here, we describe the development of an automatic breast tissue classification methodology, which can be summarized in a number of distinct steps: 1) the segmentation of the breast area into fatty versus dense mammographic tissue; 2) the extraction of morphological and texture features from the segmented breast areas; and 3) the use of a Bayesian combination of a number of classifiers. The evaluation, based on a large number of cases from two different mammographic data sets, shows a strong correlation ($\kappa = 0.81$ and 0.67 for the two data sets) between automatic and expert-based Breast Imaging Reporting and Data System mammographic density assessment.**

***Index Terms*—Breast density classification, computer-aided diagnostic systems, mammography, parenchymal patterns.**

I. Introduction

BREAST CANCER is considered a major health issue in western countries, and constitutes the most common cancer among women in the European Union. It is estimated that between 1 in 8 and 1 in 12 women will develop breast cancer during their lifetime [1], [2]. Moreover, in the European Union, as well in the United States, breast cancer remains the leading cause of death for women in their 40s [1]–[3]. However, although breast cancer incidence has increased over the past decade, breast cancer mortality has declined among women of all ages [2], [4]. This favorable trend in mortality reduction may relate to improvements made in breast cancer treatment [3] and the widespread adoption of mammography screening [4]. However, it is well known that expert radiologists can miss a significant proportion of abnormalities [5]. In addition, a large number of mammographic abnormalities turn out to be benign after biopsy [6].



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

Fig. 1. Example mammograms, where the breast density increases from (a) BIRADS I to (d) BIRADS IV [11]. See Section II for details on BIRADS categories.

Mammographic computer-aided diagnosis (CAD) systems are aimed at assisting radiologists in the evaluation of mammographic images [5], [7], and commercial systems are also available [8], [9]. Commercial and research mammographic CAD systems tend to concentrate on the detection and classification of mammographic abnormalities (e.g., microcalcifications, masses, and distortions) [10], [11]. However, recent studies have shown that the sensitivity of these CAD systems to detect mammographic abnormalities is significantly decreased as the density of the breast increases, while the specificity of the systems remains relatively constant [12]. In addition, it is well known that there is a strong positive correlation between breast tissue density in mammograms and the risk of developing breast cancer [13]–[15]. Example mammograms, covering a range of breast densities, are displayed in Fig. 1. An automatic classification of breast tissue will be beneficial, not only to decide what the density of the breast is, but also to establish an optimal strategy to follow if, for example, the user is looking for mammographic abnormalities. It should be noted that, while in this paper, we refer to *breast tissue density* and *dense tissue*, in the wider literature the following terms are also in use: parenchymal patterns, fibroglandular disk, and parenchymal density. In addition, it should be made clear that the segmentation of the dense breast tissue and mammographic risk assessment based on this dense tissue is distinct from the detection and classification of mammographic abnormalities [10], [11].

In this paper, we investigate a novel approach to automatic breast tissue classification. The initial step of the proposed methodology is the segmentation of the dense tissue region (as opposed to the region containing fatty tissue) in mammographic images, which is based on a two-class fuzzy C-means clustering approach. In the second step, morphological and texture features are extracted from the dense and fatty mammographic regions.

Manuscript received May 12, 2006; revised December 15, 2006 and April 18, 2007. This work was supported in part by the Ministerio de Educación y Ciencia (MEC) under Grant TIN2006-08035 and Grant IdIBGi-UdG 91060080. This work was supported in the form of providing the STAPLE consensus data (see Section IV-A), by Dr. S. Warfield (Computational Radiology Laboratory, Children's Hospital Boston and Harvard Medical School), whose work is supported by the National Center for Research Resources (NCRR), a component of the National Institutes of Health (NIH), under Grant R01 RR021885.

A. Oliver, J. Freixenet, and R. Martí are with the Institute of Informatics and Applications, University of Girona, 17071 Girona, Spain (e-mail: aoliver@eia.udg.es; jordif@eia.udg.es; marly@eia.udg.es).

J. Pont and E. Pérez are with the Department of Radiology, Hospital Josep Trueta, 17007 Girona, Spain.

E. R. E. Denton is with the Department of Breast Imaging, Norfolk and Norwich University Hospital, Norwich NR4 7UY, U.K.

R. Zwiggelaar is with the Department of Computer Science, University of Wales, Aberystwyth SY23 3DB, U.K. (e-mail: rrz@aber.ac.uk).

Digital Object Identifier 10.1109/TITB.2007.903514

The concatenation of the features from both regions forms a high-dimensional feature space. Dimensionality reduction and classification are achieved in the subsequent step, which relies on the combination of a number of distinct classifiers. The results provide a direct comparison between the automatic and the expert mammographic density assessment based on the Breast Imaging Reporting and Data System (BIRADS) classification protocol [11]. The evaluation is based on two independent data sets, which comprise, in total, more than a thousand mammograms. In addition, a detailed comparison with some existing methods, including results on one of the data sets, is provided.

In Section II, we provide a brief summary of a substantial number of related publications. The novelty of our approach and evaluation can be summarized by the following points: 1) the fuzzy C-means segmentation of the dense and fatty tissue regions relies on only two classes (representing dense and fatty tissue), which is in contrast with most alternative methods described in Section II that use a larger number of tissue classes representing mixtures of the two tissue types, and although no formal evaluation is provided, the segmentation provides robust and intuitively correct results; 2) at the center of our approach is the extraction of features from the segmented dense and fatty mammographic regions, which is in contrast with existing approaches in automatic mammographic risk assessment that extract information from either the full breast area (combining fatty and dense tissue information) or regions associated with the shape of the breast (again, these are likely to contain both fatty and dense tissue information); 3) the evaluation is based on two independent, publicly available, data sets: the Mammographic Image Analysis Society (MIAS) database [16] and the Digital Database of Screening Mammography (DDSM) [17], and in total, more than a thousand cases were used in the evaluation, which shows both robustness and independence of data for the developed approach; 4) the combination of all the aforementioned points provides a novel benchmark for automatic mammographic risk assessment to which newly developed techniques can be directly compared.

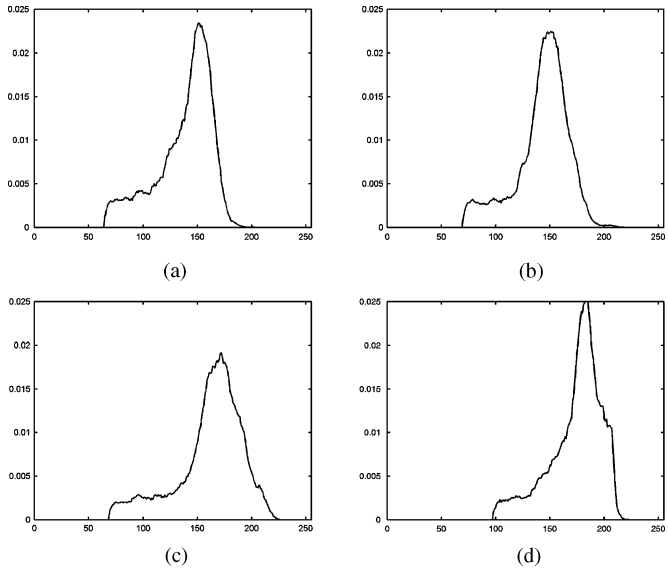
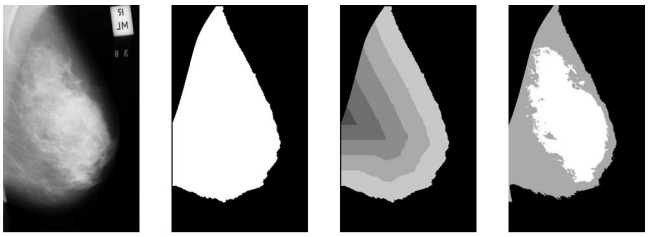


Fig. 2. Histograms associated with the mammographic images from Fig. 1, covering (a) BIRADS I to (d) BIRADS IV [11].

II. Background

Closely related to the work described here is the segmentation of the dense tissue region in mammographic images. Saha *et al*. [18] used a scale-based fuzzy connectivity method to extract dense tissue regions from mammographic images. A comparison between segmentation in craniocaudal (CC) and mediolateral-oblique (MLO) mammographic views showed a strong correlation. Ferrari *et al*. [19] used expectation maximization in combination with a minimum description length to provide the parameters for a mixture of four Gaussians. The statistical model was used to segment the fibroglandular disk, and a quantitative evaluation was provided. This work was closely related to that of Aylward *et al*. [20], who used a similar approach with a mixture of five Gaussians, although this did not exclusively concentrate on the segmentation of the fibroglandular disk. Selvan *et al*. [21] used a heuristic optimization approach to estimate model parameters for a larger number of regions. Initial segmentation results were assessed by radiologists and showed improvement when compared to alternative approaches. There is a significant volume of work on the standard mammogram form (SMF, which is also known as the h_{int} approach) by Highnam and Brady [22]. This approach relies on detailed knowledge of the mammographic system and the imaging parameters, and as such, might be less appropriate for mammograms where such information is not available (see also [23] and [24] for a detailed discussion). However, it has been shown in a number of publications by Brady's group that such an approach can be used in relation to mammographic risk assessment [25], [26], but at the same time, they have indicated that texture-based mammographic risk assessment provides improved results [27]. Closely related to the SMF-based segmentation approach is the recent work by van Engeland *et al*. [28], who used an alternative physical-image-acquisition model to estimate dense tissue volumes from full-field digital mammograms and provide a comparison with breast magnetic resonance imaging data.

The origins of breast density classification are the work of Wolfe [13], who showed the relation between mammographic parenchymal patterns and the risk of developing breast cancer, classifying the parenchymal patterns in four categories. Subsequently, Boyd *et al*. [14] showed a similar correlation between the relative area of dense tissue and mammographic risk. Since the discovery of these relationships, automated parenchymal pattern classification has been investigated [29]–[31]. These studies concentrated on the use of gray-level histograms, but recent studies have indicated that such histogram information might not be sufficient to classify mammograms according to BIRADS categories [32]–[34]. To illustrate this, Fig. 2 shows the respective histograms of the four different mammograms from Fig. 1. Note that although the mammograms belong to different classes, the four histograms are quite similar both in the mean gray-level value and the shape of the histogram. Only a small number of previous papers have suggested that



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

Fig. 3. Various strategies for dividing a mammogram (a) into regions: (b) whole breast area, (c) based on the distance between pixels and the skin line [29], [44], (d) based on clustering pixels with similar appearance.

texture representation of the breast might play a significant role. Miller and Astley [35] investigated texture-based discrimination between fatty and dense breast types. Byng *et al*. [36] used measures based on fractal dimension. Bovis and Singh [37] estimated features from the construction of spatial gray level dependency matrices. Petroudi *et al*. [27] used textons to capture the mammographic appearance within the breast area. This work was extended in the work of Petroudi and Brady [38], where the texton description was used in combination with a Markov random field approach to provide segmentation of the various tissue types, and these results formed the basis for initial (promising) classification results according to the four Wolfe classes [39]. Zwiggelaar *et al*. [40] and Zwiggelaar and Denton [41] segmented mammograms into density regions based on a set of co-occurrence matrices, and the subsequent density classification used the relative area of the density regions as the feature space. Li *et al*. [42], [43] have investigated the discriminative power of a number of texture features in classifying women with BRCA1/BRCA2 gene mutations and those at low risk. Based on regions of interest, they showed significant discriminative power for coarseness, fractal, and contrast information. The evaluation by Li *et al*. [42], [43] supports the described approach, although the texture features used here are different, these are expected to describe similar image features.

One of the main differences between some of the described approaches is the areas that are used to extract information from. To illustrate this point, a number of approaches have been depicted in Fig. 3. Bovis and Singh [37] extracted a set of features using the whole breast area, hence assuming that the breast is composed of a single texture. As shown in Fig. 3(a) and (b) (and in the mammograms shown in Fig. 1), in many cases, this is hard to justify. On the other hand, Karssemeijer [29], and subsequently, Blot and Zwiggelaar [44], divided the breast into different regions according to the distance between pixels and the skin line, as is shown in Fig. 3(a) and (c). The main idea for such an approach is the assumption that a strong correlation between tissue density and distance to the skin line will exist [42]. However, note from Fig. 3 (and again in the mammograms shown in Fig. 1) that, using such an approach, it seems that tissue with the same appearance (texture) is divided over different regions. In addition, tissues with different appearances can be merged in the same region. In contrast with these approaches, our proposal is based on the segmentation of the breast, grouping those pixels with similar tissue appearance [see Fig. 3(a) and (d)].

The classification in the mammographic risk assessment can be based on a number of categories that might not describe the same mammographic features [10], [11], [13], [14]. However, the American College of Radiology BIRADS [11] is becoming a standard on the assessment of mammographic images, which are classified in four categories according to their density (see Fig. 1 for example mammograms).

1) BIRADS I: the breast is almost entirely fatty.
2) BIRADS II: there is some fibroglandular tissue.
3) BIRADS III: the breast is heterogeneously dense.
4) BIRADS IV: the breast is extremely dense.

Although BIRADS is becoming the radiologic standard, so far, it has not been used extensively in the evaluation of automatic mammographic risk assessment approaches. Exceptions to this are the recent work of Bovis and Singh [37] and Petroudi *et al*. [27]. A full comparison between their and our results can be found in Section V.

III. Methodology

All mammograms are preprocessed to identify the breast region and remove the background, labels, and pectoral muscle areas [45]. See Fig. 3(a) and (b) for an example result of such a segmentation process. This segmentation results in a minor loss of skin-line pixels in the breast area, but those pixels are deemed not to be relevant for tissue estimation, and the relative number of potentially affected pixels is small.

A. Finding Regions With Similar Tissue

Gray-level information in combination with the fuzzy C-means clustering approach [46] is used to group pixels into two separate categories: fatty and dense tissues. Prior to this, and with the aim of avoiding effects from microtexture that could appear in some regions, the breast region is smoothed by using a median filter of size 5×5. From our experiments, this filter size is a good compromise between noise reduction and texture preservation of mammographic tissue.

Fuzzy C-means is an extension of the well-known k-means algorithm [47]. The main difference is that fuzzy C-means allows each pattern of the image to be associated with every cluster using a fuzzy membership function (in *k*-means, each pattern belongs to one and only one cluster). In our implementation, the function criterion minimized by the algorithm is defined by

$$e^2(\Xi, U) = \sum_{n=1}^{N} \sum_{t=1}^{T} u_{nt} ||p_n - c_t||^2 \tag{1}$$

where Ξ is the partition of the image, U is the membership matrix: u_{nt} represents the membership of the pattern p_n to belong to cluster t, which is centered at

$$c_t = \frac{\sum_{n=1}^{N} u_{nt} p_n}{\sum_{n=1}^{N} u_{nt}} \tag{2}$$

N is the number of patterns in the whole image (the number of pixels) and T the number of clusters, which has to be known *a priori* (here $T = 2$ representing fatty and dense tissue).

When using partitional clustering algorithms, like fuzzy C-means, the placement of the initial seed points is one of the central issues in the variation of segmentation results [48]. Despite their importance, seeds for these algorithms are usually initialized randomly. The two classes in our approach were initialized with the gray-level values that represent 15% and 85% of the accumulative histogram of the breast pixels of each mammogram (representing fatty and dense tissue, respectively). Although these values were determined empirically, the obtained segmentations do not critically depend on them. Moreover, some mammograms do not have clearly determined dense and fatty components. In these cases, the result of the segmentation is one cluster grouping the breast tissue and the other cluster grouping regions with less compressed tissue (an elongated region, like a ribbon, following the skin line). In these cases, the breast texture information is in the breast tissue cluster, while the ribbon does not provide significant information to the system.

B. Extracted Features

The result of the fuzzy C-means algorithm is the division of the breast into (only) two clusters. Subsequently, a set of features for both classes can be directly extracted from the original images (no preprocessing/filtering was applied). Here, we used a set of morphological and texture features. As morphological features, the relative area and the first four histogram moments for both clusters were calculated. Note that the four moments of the histogram are related to the mean intensity, the standard deviation, the skewness, and the kurtosis of each cluster.

A set of features derived from co-occurrence matrices [49] were used as texture features. Co-occurrence matrices are essentially two-dimensional histograms of the occurrence of pairs of gray levels for a given displacement vector. Formally, the co-occurrence of gray levels can be specified as a matrix of relative frequencies P_{ij}, in which two pixels separated by a distance d and angle θ have gray levels i and j. Here, we use four different directions: 0°, 45°, 90°, and 135°, and three distances equal to 1, 5, and 9 pixels. Note that these values were determined empirically and are related to the scale of the texture features found in mammographic images. The full co-occurrence matrices are generally not used as features (mainly due to their high dimensionality and potential sparseness), but instead, a large number of features derived from such matrices have been proposed [49]. For each co-occurrence matrix, the following features were used: contrast, energy, entropy, correlation, sum average, sum entropy, difference average, difference entropy, and homogeneity features.

As each of these features is extracted from each class, we deal with 226 features in total, 10 from morphological characteristics and 216 from the texture information.

C. Classification

The classification of mammograms according to BIRADS categories was performed in three different ways: by using a k-nearest neighbors (kNN) classifier, a decision tree classifier, and a Bayes classifier based on the combination of the first two classifiers. Both the kNN and the decision tree classifiers can naturally deal with multiclass data, which is less the case for some of the more advanced classifiers (although it should be mentioned that it is possible to combine the results of a combinatoric set of such classifiers to simulate multiclass results) [50], [51].

A combination of classifiers is investigated as it has been shown in an initial study that kNN and ID3 classifiers (which are the basis for the classifiers used in this paper) provide complementary information [32], [34]. Based on this, we are assuming independence of classifiers and have used a Bayesian approach to combine the kNN and decision tree results. Such an approach is further supported by work on ensemble classifiers [52], [53], which has shown benefits in robustness and classification performance for such ensembles when compared to single nonlinear state-of-the-art classifiers (see also the discussion presented in [54] and [55]). It should be noted that boosting, which is an ensemble classifier methodology, is being used in the decision tree classifier approach described in Section III-C.2.

1) kNN Classification: The kNN classifier [47] consists of the assignment of an unclassified vector using the closest k vectors found in the training set. Here, the Euclidean distance is used. Due to the fact that kNN is based on distances between sample points in the feature space, features need to be normalized to prevent some features being more strongly weighted than others. Hence, all features have been normalized to unit variance and zero mean. Moreover, kNN presents another inherent problem, which is the uniform weighting of features regardless of their discriminant power. In order to solve this problem we have included a feature-selection step that automatically selects the set of the most discriminant features. Here, we have used the sequential forward selection (SFS) algorithm [56], which is a widely known technique that selects a local optimum solution in a computationally attractive way.

2) Decision Tree Classification: The second classifier used is a decision tree. A decision tree recursively subdivides regions in the feature space into different subspaces, using different thresholds in each dimension to maximize class discrimination. Ideally, for a given subspace, the process stops when it only contains patterns of one class. However, in practice, sometimes it is not possible or is computationally prohibitive to use such a stopping criterion, and the algorithm stops when most of the patterns belong to the same region. Here, we have used the C4.5 decision tree [57], which is an extension of the ID3 decision tree [58] and deals naturally with continuous data. In order to obtain a more robust classifier, the boosting procedure described in [59] is used.

3) Combined Bayesian Classification: The third classifier is based on a combination of the two classifiers described earlier and uses a Bayesian estimation approach [47]. When a new case is studied, it is classified according to the classic Bayes equation:

$$P(x \in B_c | A(x)) = \frac{P(A(x)|x \in B_c)P(B_c)}{\sum_{l=1,\ldots,4} P(A(x)|x \in B_l)P(B_l)}. \quad (3)$$

Translating this formula into words, we consider the probability of a mammogram x, with set of features $A(x)$, to belong to the class B_c as the posterior probability. The prior is the probability of the mammogram to belong to a class before any observation of the mammogram. If there were the same number of cases for each class, the prior would be constant (for four categories, as is the case for BIRADS classification, and hence, $l = 1, \ldots, 4$, the constant value would be 0.25). Here, we used as the prior probability the number of cases that exist in the database for each class, divided by the total number of cases. The likelihood estimation is calculated by using a nonparametric estimation, which is explained in the next paragraph. Finally, the evidence includes a normalization factor, needed to ensure that the sum of posteriors probabilities for each class is equal to 1.

Combining the SFS + kNN and C4.5 classifiers is achieved by a soft-assign approach where binary (or discrete) classification results are transformed into continuous values that depict class membership. For the SFS+kNN classifier, the membership value of a class is proportional to the number of neighbors belonging to this class. The membership value for each class B_c will be the sum of the inverse Euclidean distances between the k neighboring patterns belonging to that class and the unclassified pattern:

$$P_{\mathrm{kNN}}(A(x)|x \in B_c) = \sum_{\xi \in \mathrm{kNN} \wedge B_c} \frac{1}{1 + \mathrm{dist}(A(x), A(\xi))}. \quad (4)$$

Note that with this definition, a final normalization to one over all the membership values is required. On the other hand, in the boosted C4.5 decision tree, a new pattern is classified by using the vote of the different classifiers weighted by their accuracy. Thus, in order to achieve a membership for each class, instead of considering the voting criteria, we take into account the result of each classifier. Adding all the results for the same class and normalizing all the results, the membership for each class is obtained.

D. Evaluation

The evaluation of the automatic and manual density classification is presented in the form of confusion matrices [47]. For each confusion matrix, we include the kappa (κ) coefficient [60]. This is used as a means of estimating agreement in categorical data, and is computed as:

$$\kappa = \frac{P(D) - P(E)}{1 - P(E)} \quad (5)$$

where $P(D)$ is the proportion of times the model value was equal to the actual value (the diagonal terms) and $P(E)$ is the expected proportion by chance. A κ coefficient equal to 1 means a statistically perfect model whereas a value equal to 0 means every model value was different from the actual value. Table I shows a commonly used interpretation of the various κ values [61]. In addition, the classification results that will be discussed are given as a correct classification percentage (CCP in Tables III–IV). For the overall classification results, this is determined by the sum of the diagonal elements of the confusion matrices divided by the total number of mammograms in the used dataset. For individual BIRADS classification, this reduces to the number of correctly classified mammograms divided by the total number of mammograms for that density class.

TABLE I
COMMON INTERPRETATION OF THE VARIOUS κ VALUES [61]

κ	Agreement
< 0	Poor
$[0, 0.20]$	Slight
$[0.21, 0.40]$	Fair
$[0.41, 0.60]$	Moderate
$[0.61, 0.80]$	Substantial
$[0.81, 1.00]$	Almost Perfect

An alternative approach to summarize the evaluation results would be the use of receiver operator characteristic (ROC) curves [62]. However, in the described evaluation, this is less appropriate because, in general, ROC analysis assumes binary classification results. Such binary results can be enforced by considering each density class on its own after which the results could be combined. This is not a true representation of the four-density-class problem, as described here, and as such, ROC analysis will not be used in the evaluation.

IV. RESULTS

In order to test the proposed method, two public and widely known databases were used: the MIAS database [16] and the DDSM database [17]. While the latter has its density classified using BIRADS categories, the former only uses three classes. Three mammography experts have classified all the MIAS mammograms according to the BIRADS lexicon.

A. MIAS Database

The method was applied to the whole set of 322 mammograms that form the MIAS database [16]. This database is composed of MLO left and right mammograms from 161 women. The spatial resolution of the images is $50\ \mu\mathrm{m} \times 50\ \mu\mathrm{m}$ and quantized to 8 bits with a linear optical density in the range 0–3.2.

Three expert mammographic readers classified all the images in the MIAS database according to the BIRADS categories (the correlation between the original triple MIAS and BIRADS classification is discussed in [63]). In screening mammography, it is common to obtain expert agreement; here, a similar approach is used, and consensus between the individual expert classification is used. Table II shows the confusion matrices for the classification of the three radiologists and the consensus opinion. This consensus is determined by selecting as the final class, the class where two or three radiologists agreed (majority vote). If the three experts classified the mammogram in different classes, the median value is selected as the consensus opinion. The results in Table II show divergence in the opinion of the radiologists, directly indicating the difficulty of the problem we are dealing with, and indicate the need to remove interobserver (interoperator) variability through the development of automatic methods.

Using the κ-values, the agreement of experts A and C with the consensus opinion fall in the *substantial* category, while the agreement of expert B and the consensus opinion belongs to the *almost perfect* category (i.e., the classification by expert B

TABLE II
CONFUSION MATRICES FOR THREE EXPERT RADIOLOGISTS AND THEIR CONSENSUS OPINION

Expert A ($\kappa = 0.70$)

Consensus \ Expert A	B-I	B-II	B-III	B-IV
B-I	85	2	0	0
B-II	43	60	0	0
B-III	1	17	70	7
B-IV	0	0	0	37

Expert B ($\kappa = 0.85$)

Consensus \ Expert B	B-I	B-II	B-III	B-IV
B-I	85	2	0	0
B-II	1	93	9	0
B-III	0	17	72	6
B-IV	0	0	0	37

Expert C ($\kappa = 0.61$)

Consensus \ Expert C	B-I	B-II	B-III	B-IV
B-I	59	28	0	0
B-II	0	58	45	0
B-III	0	0	88	7
B-IV	0	0	10	27

is almost equal to the consensus). Compared to the consensus, expert C shows a slight bias toward the higher BIRADS classes than do the other two experts, while expert A shows a slight bias toward the lower BIRADS classes.

Instead of using the majority vote to provide the consensus classification, it is possible to use an expectation maximization approach like STAPLE [64]. In this case, STAPLE produced a consensus that was very close to the majority vote results, with only two mammograms being classed differently. This has minimal effects on the results: the maximum difference on the overall classification results being $\pm 0.3\%$, while for the individual BIRADS classes, this increases to $\pm 1.1\%$ (and here, positive changes for one BIRADS class are matched by negative changes for one of the other BIRADS classes). For the remainder of the paper, we have used the majority vote results as the consensus-classification results.

In order to test the proposed method, we performed two experiments related to the experts classification. Firstly, by training the classifiers based on the ground truth as provided by the individual experts, we can evaluate the correlation between the methods and each radiologist. The second experiment was performed by training the classifier using the consensus between all three experts as the ground truth. In this case, we would expect a better agreement as the interobserver variability is minimized.

1) Results Based on Individual Manual Classification: Initial experiments consist of the evaluation of the proposed method using the individual expert classifications independently. We used a leave-one-woman-out methodology, i.e., the left and right mammograms of a woman are analyzed by a classifier trained using the mammograms of all other women in the database. The leave-one-woman-out methodology is used to avoid bias as the left and right mammograms of a woman are expected to have similar internal morphology [65]. The confusion matrices for the three classifiers: the SFS + kNN, C4.5, and Bayesian approaches are shown in Table III, where each row corresponds to results based on the manual classification by an individual radiologist. In this paper, a value of $k = 7$ was used for kNN. Other odd values ranging from 5 to 15 were tested, and gave similar results.

For expert A, the SFS+kNN correctly classifies about 78% $(252/322)$ of the mammograms, while the C4.5 decision tree achieves 74% (239/322) of correct classification. SFS+kNN clearly outperforms C4.5 when classifying mammograms belonging to BIRADS II, while for the rest of BIRADS, the performance is quite similar. On the other hand, C4.5 tends to classify the mammograms according to its own or its neighboring BIRADS classification, while SFS+kNN shows a larger dispersion. The κ coefficient also reflects that SFS+kNN has better performances than that of C4.5, with values equal to 0.70 and 0.64, respectively. Note that results for both classifiers belong to the *substantial* category according to Table I.

The results obtained by the Bayesian classifier are shown in Table III(C)). This classifier shows an increase in the overall performance when compared to the individual classifiers, reaching 83% (266/322) correct classification. This is an increase of 5% and 9% when compared to SFS+kNN and C4.5, respectively. When considering the individual BIRADS classes, the percentage of correct classification for BIRADS I is around 91% (118/129), while in the other cases, the percentages are 76% (60/78) for BIRADS II, 76% (53/70) for BIRADS III, and 80% (35/44) for BIRADS IV. Note that using the Bayesian classifier, κ is increased to 0.76.

The results obtained for expert B are slightly decreased with respect to those obtained for expert A. Specifically, 74% (238/322) of the mammograms were correctly classified by using the SFS+kNN classifier, while the C4.5 results remained at 67% $(217/322)$. The better results for the SFS+kNN classifier are independent of the BIRADS classes, except for the BIRADS IV class, in which C4.5 clearly outperforms SFS+kNN. The results obtained by the Bayes classifier shows an increase in the performance of 6% and 13% when compared to SFS+kNN and C4.5, respectively, obtaining an overall performance of 80% (257/322). When considering the individual BIRADS classes, the correct classification percentage for BIRADS I is around 91% (78/86), while for the other cases, the percentages are 83% (93/112) for BIRADS II, 68% (55/81) for BIRADS III, and 74% (32/43) for BIRADS IV. The κ-value is equal to 0.73.

The last row of Table III shows the results obtained for expert C. The performance of the classifiers is similar to that obtained by using the ground truth of expert B. The SFS+kNN classifier obtained 74% (239/322) correct classification, while C4.5 obtained 72% (231/322). Using the Bayes classifier, 82% (263/322) of the mammograms were correctly classified. In summary, 86% (51/59) correct classification for BIRADS I, 74% (64/86) for BIRADS II, 85% (122/143) for BIRADS III, and 78% (26/34) for BIRADS IV. The κ-value is 0.73.

In summary, the best classification rates are obtained using the Bayesian combination. For each individual expert, 83%, 80%, and 82% correct classification are obtained, respectively.

In line with other publications [27], [37], we can reduce the four-class classification problem to the following two-class problem: $\{$BIRADS I and II$\}$ versus $\{$BIRADS III and IV$\}$, or

TABLE III
CONFUSION MATRICES FOR MIAS CLASSIFICATION FOR INDIVIDUAL EXPERTS CLASSIFICATION: (A) SFS+KNN, (B) C4.5 DECISION TREE, AND (C) BAYESIAN CLASSIFIERS

Expert A

SFS+kNN ($\kappa = 0.70$; $CCP = 78\%$)

	B-I	B-II	B-III	B-IV
B-I	113	10	5	1
B-II	8	59	9	3
B-III	4	13	46	7
B-IV	1	3	6	34

C4.5 ($\kappa = 0.64$; $CCP = 74\%$)

	B-I	B-II	B-III	B-IV
B-I	114	12	2	1
B-II	18	47	12	2
B-III	2	11	48	9
B-IV	0	1	13	30

Bayesian ($\kappa = 0.76$; $CCP = 83\%$)

	B-I	B-II	B-III	B-IV
B-I	118	6	5	0
B-II	7	60	10	2
B-III	0	6	53	11
B-IV	0	2	7	35

Expert B

SFS+kNN ($\kappa = 0.64$; $CCP = 74\%$)

	B-I	B-II	B-III	B-IV
B-I	75	8	2	1
B-II	7	85	16	4
B-III	1	20	55	5
B-IV	2	7	11	23

C4.5 ($\kappa = 0.55$; $CCP = 67\%$)

	B-I	B-II	B-III	B-IV
B-I	69	15	2	0
B-II	13	73	22	4
B-III	1	27	46	7
B-IV	0	1	13	29

Bayesian ($\kappa = 0.73$; $CCP = 80\%$)

	B-I	B-II	B-III	B-IV
B-I	78	6	2	0
B-II	10	93	8	1
B-III	0	16	55	10
B-IV	0	1	10	32

Expert C

SFS+kNN ($\kappa = 0.63$; $CCP = 74\%$)

	B-I	B-II	B-III	B-IV
B-I	50	5	1	3
B-II	13	53	19	1
B-III	0	21	115	7
B-IV	3	3	7	21

(A)

C4.5 ($\kappa = 0.58$; $CCP = 72\%$)

	B-I	B-II	B-III	B-IV
B-I	43	14	0	2
B-II	15	49	22	0
B-III	2	15	119	7
B-IV	1	0	13	20

(B)

Bayesian ($\kappa = 0.73$; $CCP = 82\%$)

	B-I	B-II	B-III	B-IV
B-I	51	5	1	2
B-II	9	64	12	1
B-III	1	16	122	4
B-IV	0	2	6	26

(C)

TABLE IV
CONFUSION MATRICES FOR MIAS CLASSIFICATION USING THE CONSENSUS CLASSIFICATION: (A) SFS+KNN, (B) C4.5 DECISION TREE, AND (C) BAYESIAN CLASSIFIERS

Consensus

SFS+kNN ($\kappa = 0.68$; $CCP = 77\%$)

	B-I	B-II	B-III	B-IV
B-I	70	13	1	3
B-II	9	80	13	1
B-III	1	17	73	4
B-IV	3	2	8	24

(A)

C4.5 ($\kappa = 0.61$; $CCP = 72\%$)

	B-I	B-II	B-III	B-IV
B-I	72	13	1	1
B-II	13	68	20	2
B-III	0	21	68	6
B-IV	0	2	11	24

(B)

Bayesian ($\kappa = 0.81$; $CCP = 86\%$)

	B-I	B-II	B-III	B-IV
B-I	79	1	3	4
B-II	3	86	6	8
B-III	0	2	85	8
B-IV	0	6	4	27

(C)

in words, low density (low risk) versus high density (high risk) classification, which, from a mammographic risk assessment point of view, might be more appropriate than the four-class division. For expert A, the percentage of correct classification is about 92% for the three classifiers and low breast densities, while for dense breasts, the percentage is 82%, 88%, and 93% for the SFS+kNN, C4.5, and the Bayesian combination, respectively. In contrast, for expert B, the correct classification percentage for low-density breasts is around 88% for the single classifiers and 94% for the combination, while for high-density breasts, it is reduced to 76% for each classifier, and 86% for their combination. On the other hand, using expert C, the correct classification percentage for low-density breasts is 83% for the single classifiers and 89% for the combination, while for high-density breasts, the SFS+kNN obtains 85%, and the other classifiers 89%.

For the two-class approach, in summary, the results are 92%, 91%, and 89% of correct classification for experts A, B, and C, respectively.

2) Results Based on Consensus Manual Classification: Table IV shows results based on a leave-one-woman-out methodology (see before for a more detailed discussion) for the classification of the whole MIAS database according to the consensus ground truth. The performance of the individual classifiers is 77% (247/322) correct classification for SFS+kNN and 72% (232/322) for C4.5. These are intermediate values between expert A and both experts B and C. However, the Bayesian combination of the classifiers results in a substantial improvement and 86% (277/322) correct classification is achieved, which gives a better performance compared to those obtained by the individual experts without consensus. This result is confirmed by $\kappa = 0.81$, which belongs just to the *almost perfect* category. Examining each class, BIRADS I reached 91% (79/87) correct classification, BIRADS II 84% (86/103), BIRADS III 89% (85/95), and BIRADS IV 73% (27/37).

Using the two-class classification, low-density mammograms are 89% correctly classified, while high-density ones reach 94%, resulting in an overall classification equal to 91%.

B. DDSM Database

The developed methodology was also evaluated on a set of 831 mammograms taken from the DDSM [17], with the main objective to demonstrate the robustness of our proposal on a different and larger data set. Similarly to the MIAS database, DDSM provides for each mammogram additional information including the density of the breast. In contrast to MIAS, this information is determined using the BIRADS categories.

The number of mammograms belonging to each category is 106 (13%), 336 (40%), 255 (31%), and 134 (16%) for BIRADS I–IV, respectively. These proportions are consistent with the numbers reported by ongoing screening programs. As shown in the work of Lehman *et al.* [66], where a population of $46\,340$ women was studied: 13.6% were BIRADS I, 50.9% BIRADS

TABLE V
CONFUSION MATRICES FOR DDSM CLASSIFICATION: (A) SFS+KNN, (B) C4.5 DECISION TREE, AND (C) BAYESIAN CLASSIFIERS

SFS+kNN ($\kappa = 0.56$; $CCP = 70\%$)

Truth	B-I	B-II	B-III	B-IV
B-I	54	40	12	0
B-II	44	266	25	1
B-III	9	60	177	9
B-IV	0	21	30	83

(A)

C4.5 ($\kappa = 0.59$; $CCP = 72\%$)

	B-I	B-II	B-III	B-IV
B-I	51	30	25	0
B-II	22	279	35	0
B-III	16	59	178	2
B-IV	8	14	25	87

(B)

Bayesian ($\kappa = 0.67$; $CCP = 77\%$)

	B-I	B-II	B-III	B-IV
B-I	58	25	23	0
B-II	15	295	26	0
B-III	12	46	196	1
B-IV	5	18	18	93

(C)

II, 30.1% BIRADS III, and 5.5% BIRADS IV. Although these percentages vary with the age of the women, classes II and III tend to be larger than classes I and IV [67]–[69].

The DDSM database provides four mammograms, comprising left and right MLO and left and right CC views, for most women. To avoid bias we selected only the right MLO mammogram for each woman. This way, the leave-one-woman-out method used for evaluating the system in the previous sections is now reduced to the typical leave-one-image-out evaluation methodology.

Using this evaluation strategy, Table V shows the results obtained with the classifiers. These results show a slightly reduced performance when compared to the MIAS database results (see Tables III and IV). To be specific, the performance obtained by the classifiers is 70% (580/831), 72% (595/831), and 77% (642/831) for SFS+kNN, C4.5, and Bayesian combination, respectively. Note that by using this database, the performance using C4.5 is better than that by using SFS+kNN. This can be due to the use of more mammograms and a different distribution over the BIRADS classes in the training set. The κ-value, equal to 0.67, indicates a *substantial* correlation between the manual and the automatic Bayesian classification.

Examining each class alone, BIRADS I reached 55% (58/106) correct classification, BIRADS II 88% (295/336), BIRADS III 77% (196/255), and BIRADS IV 69% (93/134). In contrast to the MIAS database, BIRADS I shows the worst results, while BIRADS II shows the best. We believe that this result is due to the fact that, in the DDSM database, mammograms belonging to BIRADS I have tissue very similar to those belonging to BIRADS II. Related to the classification of dense mammograms, the ones belonging to BIRADS III are better classified than the ones belonging to BIRADS IV. Moreover, only one mammogram not belonging to BIRADS IV is incorrectly classified as belonging to this class.

Using the low-/high-density division, low-density mammograms are 89% correctly classified, while high-density ones reach 79%. It should be clear that compared to the MIAS consensus results, the performance is mainly reduced on the high-density mammograms, while a similar classification is obtained for the low-density ones. The overall performance for the Bayesian two-class classifier is equal to 84%.

V. DISCUSSION

In order to perform a quantitative comparison, we implemented the algorithm described in [29] (see also Section II) using the MIAS database for evaluation. Two different experiments were conducted related to the features used. Firstly, the features described in [29], except for the pectoral muscle gray-level information, were used in combination with a kNN approach. The results, shown in the top confusion matrix of Table VI, indicate a correct classification level equal to 66% (213/322), which is in close agreement with the values quoted in [29] (it should be noted that they showed improved results on more recent mammograms). Secondly, the features described in Section III-B were used in combination with the Bayesian classifier (see Section III-C), and the results, shown at the bottom confusion matrix of Table VI, indicate an improvement to 75% (242/322) correct classification. The last result illustrates the benefits that the used feature set has over alternative descriptors.

TABLE VI
CONFUSION MATRICES FOR MIAS CLASSIFICATION USING THE CONSENSUS CLASSIFICATION: (A) USING THE SEGMENTATION AND THE FEATURES DESCRIBED IN [29], (B) USING THE SEGMENTATION DESCRIBED IN [29] BUT WITH THE FEATURES USED IN OUR WORK

(A) ($\kappa = 0.53$; $CCP = 66\%$)

Consensus	B-I	B-II	B-III	B-IV
B-I	70	10	5	2
B-II	9	62	30	2
B-III	1	21	64	9
B-IV	3	5	12	17

(B) ($\kappa = 0.65$; $CCP = 75\%$)

Consensus	B-I	B-II	B-III	B-IV
B-I	77	9	1	0
B-II	11	77	14	1
B-III	0	25	66	4
B-IV	1	5	9	22

The results are summarized in Table VII, which also includes the results obtained with the proposed approach, reaching 86% correct classification using the same database. This indicates that the improvements over existing techniques can be associated with two aspects, which are the features that are being used and the fact that these features are extracted from the fatty and dense tissue regions.

In the literature, only the works of Bovis and Singh [37] and Petroudi *et al.* [27] have classified breast tissue according to BIRADS categories. While Bovis and Singh reached 71% of correctly classified mammograms, Petroudi *et al.* achieved an overall correct classification of 76%. Table VII summarizes in more detail the results they obtained. It can be seen that Petroudi *et al.* obtained similar results to our MIAS-database-based evaluation, but with significant lesser results on BIRADS II and III, and hence, on the overall classification (column $Total_4$). Moreover, the table shows that Bovis has lower four-class results on a smaller DDSM dataset, but higher overall low/high classification (column $Total_2$). Note, however, that a direct comparison is difficult as both these publications have used different

TABLE VII
SUMMARY OF RESULTS AND COMPARISON WITH EXISTING WORK USING: (A) THE SEGMENTATION APPROACH AND THE FEATURES DESCRIBED IN [29], (B) THE SEGMENTATION APPROACH DESCRIBED IN [29] BUT WITH THE FEATURES AS USED IN OUR WORK, AND (C) THE DEVELOPED APPROACH

	BIRADS CDID (%)				Correct Classification Percentage (CCP)							
	B-I	B-II	B-III	B-IV	B-I	B-II	B-III	B-IV	$Total_4$	Low	High	$Total_2$
(A) - MIAS	27	32	30	11	80	60	67	46	66	79	77	79
(B) - MIAS	27	32	30	11	89	75	69	59	75	92	77	85
(C) - MIAS	27	32	30	11	91	84	89	73	86	89	94	91
(C) - DDSM	13	40	31	16	55	88	77	69	77	89	79	84
Bovis [37]	20	21	25	33					71			97
Petroudi [27]					91	64	70	78	76	91	94	

datasets. Bovis and Singh used 377 DDSM MLO images (probably different from the ones used in our work), while Petroudi *et al*. used 132 local (nonpublicly available) CC/MLO images. It is likely that the distribution over the various BIRADS categories is different in each experiment, and as such, could influence the results in that a dataset with a distribution skewed toward BIRADS classes I and IV can be expected to show better results than that of a dataset with a distribution with a higher proportion of II and III category images. In our experiments a similar behavior could be seen in the results obtained using the MIAS database and expert A, who, in comparison with experts B and C, used a high percentage of BIRADS I classifications.

The strength of the Bayesian classifier might be partially explained by the features that were mainly used by the individual classifiers. The SFS stage of the SFS+kNN classifier has a strong tendency to select texture features independently of the distance used for the co-occurrence matrices, while most of the selected features for the C4.5 classifiers are related to the statistics obtained using a distance equal to 9 for the co-occurrence matrices.

A direct comparison with alternative mammographic dense tissue segmentation techniques [18]–[22], [25], [26], [28] is seen to be outside the scope of this paper as this segmentation forms only one of the steps in the described approach. However, in a recent study [70], we have shown that mammographic risk assessment based on identical features extracted after using various mammographic segmentation techniques resulted in significant differences, with classification based on two-class segmentation (as the fuzzy C-means approach used here) providing superior results. However, all the various two-class segmentation approaches resulted in similar classification results, which might indicate that this is not a limiting factor in the overall system.

Future work will focus on two areas. Firstly, a larger case study will be performed in a clinical environment. Secondly, it is our aim to investigate the behavior of the method for full-field digital mammograms, as the final goal of the methodology is integration in a CAD tool.

VI. CONCLUSION

To summarize the developed method, the initial step, based on gray-level information, segments the breast area. Subsequently, the fuzzy C-means algorithm is used to segment fatty versus dense tissue types in the mammograms. For each tissue region, morphological and texture features are extracted to characterize the breast tissue. Finally, using a Bayesian approach and obtaining the likelihood estimation by combining both SFS+kNN and C4.5 classifier results, the mammograms are classified according to BIRADS categories. It should be noted that, to avoid bias, we have adopted a leave-one-woman-out methodology.

Summarizing the results, we obtained for the MIAS database and individual experts 83%, 80%, and 82% correct classification, which increased to 86% when the classifiers were based on the consensus ground truth. On the other hand, results based on the DDSM database (a set of 831 mammograms) showed a performance of 77% correct classification.

ACKNOWLEDGMENT

The authors would also like to thank the reviewers for their comments that improved the content and readability of the paper considerably.

REFERENCES

[1] Eurostat, "Health statistics: Atlas on mortality in the European Union," Office for Official Publications of the European Union, 2002.
[2] F. Bray, P. McCarron, and D. M. Parkin, "The changing global patterns of female breast cancer incidence and mortality," *Breast Cancer Res.*, vol. 6, pp. 229–239, 2004.
[3] S. Buseman, J. Mouchawar, N. Calonge, and T. Byers, "Mammography screening matters for young women with breast carcinoma," *Cancer*, vol. 97, no. 2, pp. 352–358, 2003.
[4] E. A. Sickles, "Breast cancer screening outcomes in women ages 40–49: Clinical experience with service screening using modern mammography," *J. Nat. Cancer Inst.: Monographs*, vol. 22, pp. 99–104, 1997.
[5] R. L. Birdwell, D. M. Ikeda, K. D. O'Shaughnessy, and E. A. Sickles, "Mammographic characteristics of 115 missed cancers later detected with screening mammography and the potential utility of computer-aided detection," *Radiology*, vol. 219, pp. 192–202, 2001.
[6] F. M. Hall, J. M. Storella, D. Z. Siverstond, and G. Wyshak, "Nonpalpable breast lesions: Recommendations for biopsy based on suspicion of carcinoma at mammography," *Radiology*, vol. 167, pp. 353–358, 1988.
[7] T. W. Freer and M. J. Ulissey, "Screening mammography with computer-aided detection: Prospective study of 12860 patients in a community breast center," *Radiology*, vol. 220, pp. 781–786, 2001.
[8] R2 ImageChecker. (2007 Jan. 1) [Online]. Available: http://www.r2tech.com.
[9] iCAD Second Look. (2007 Jan. 1) [Online]. Available: http://www.icadmed.com.
[10] L. Tabar, T. Tot, and P. B. Dean, *Breast Cancer—The Art and Science of Early Detection With Mammography*. Stuttgart, Germany: Georg Thieme Verlag, 2005.
[11] American College of Radiology, *Illustrated Breast Imaging Reporting and Data System BIRADS*, 3rd ed. Philadelphia, PA: Amer. College of Radiol, 1998.
[12] W. T. Ho and P. W. T. Lam, "Clinical performance of computer-assisted detection (CAD) system in detecting carcinoma in breasts of different densities," *Clin. Radiol.*, vol. 58, pp. 133–136, 2003.
[13] J. N. Wolfe, "Risk for breast cancer development determined by mammographic parenchymal pattern," *Cancer*, vol. 37, pp. 2486–2492, 1976.

[14] N. F. Boyd, J. W. Byng, R. A. Jong, E. K. Fishell, L. E. Little, A. B. Miller, G. A. Lockwood, D. L. Tritchler, and M. J. Yaffe, "Quantitative classification of mammographic densities and breast cancer risk: Results from the Canadian national breast screening study," *J. Nat. Cancer Inst.*, vol. 87, pp. 670–675, 1995.

[15] V. A. McCormack and I. dos Santos Silva, "Breast density and parenchymal patterns as markers of breast cancer risk: A meta-analysis," *Cancer Epidemiol. Biomarkers Prev.*, vol. 15, pp. 1159–1169, 2006.

[16] J. Suckling, J. Parker, D. R. Dance, S. M. Astley, I. Hutt, C. R. M. Boggis, I. Ricketts, E. Stamatakis, N. Cerneaz, S. L. Kok, P. Taylor, D. Betal, and J. Savage, "The Mammographic Image Analysis Society digital mammogram database," in *Proc. Int. Workshop Dig. Mammography*, 1994, pp. 211–221.

[17] M. Heath, K. Bowyer, D. Kopans, R. Moore, and P. J. Kegelmeyer, "The digital database for screening mammography," in *Proc. Int. Workshop Dig. Mammography*, 2000, pp. 212–218.

[18] P. K. Saha, J. K. Udupa, E. F. Conant, P. Chakraborty, and D. Sullivan, "Breast tissue density quantification via digitized mammograms," *IEEE Trans. Med. Imag.*, vol. 20, no. 8, pp. 792–803, Aug. 2001.

[19] R. J. Ferrari, R. M. Rangayyan, R. A. Borges, and A. F. Frere, "Segmentation of the fibro-glandular disc in mammograms via Gaussian mixture modelling," *Med. Biol. Eng. Comput.*, vol. 42, pp. 378–387, 2004.

[20] S. R. Aylward, B. H. Hemminger, and E. D. Pisano, "Mixture modelling for digital mammogram display and analysis," in *Proc. Int. Workshop Dig. Mammography*, 1998, pp. 305–312.

[21] S. E. Selvan, C. C. Xavier, N. Karssemeijer, J. Sequeira, R. A. Cherian, and B. Y. Dhala, "Parameter estimation in stochastic mammogram model by heuristic optimization techniques," *IEEE Trans. Inf. Technol. Biomed.*, vol. 10, no. 4, pp. 685–695, Oct. 2006.

[22] R. Highnam and M. Brady, *Mammographic Image Analysis*. Dordrecht, The Netherlands: Kluwer, 1999.

[23] L. Blot and R. Zwiggelaar, "A volumetric approach to risk assessment in mammography: A feasibility study," *Phys. Med. Biol.*, vol. 50, pp. 695–708, 2005.

[24] R. Highnam, X. Pan, R. Warren, M. Jeffreys, G. D. Smith, and M. Brady, "Breast composition measurements using retrospective standard mammogram form (SMF)," *Phys. Med. Biol.*, vol. 51, pp. 2695–2713, 2006.

[25] K. Marias, S. Petroudi, R. English, R. Adams, and M. Brady, "Subjective and computer-based characterisation of mammographic patterns," in *Proc. Int. Workshop Dig. Mammography*, 2002, pp. 552–556.

[26] S. Petroudi, K. Marias, R. English, R. Adams, and M. Brady, "Classification of mammographic patterns using area measurements and the standard mammogram form (SMF)," in *Proc. Med. Image Understanding Anal. Conf.*, 2002, pp. 197–200.

[27] S. Petroudi, T. Kadir, and M. Brady, "Automatic classification of mammographic parenchymal patterns: A statistical approach," in *Proc. IEEE Conf. Eng. Med. Biol. Soc.*, 2003, vol. 1, pp. 798–801.

[28] S. van Engeland, P. R. Snoeren, H. Huisman, C. Boetes, and N. Karssemeijer, "Volumetric breast density estimation from full-field digital mammograms," *IEEE Trans. Med. Imag.*, vol. 25, no. 3, pp. 273–282, Mar. 2006.

[29] N. Karssemeijer, "Automated classification of parenchymal patterns in mammograms," *Phys. Med. Biol.*, vol. 43, pp. 365–378, 1998.

[30] 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," *Med. Phys.*, vol. 28, no. 6, pp. 1056–1069, 2001.

[31] K. E. Martin, M. A. Helvie, C. Zhou, M. A. Roubidoux, J. E. Bailey, C. Paramagul, C. E. Blane, K. A. Klein, S. S. Sonnad, and H. P. Chan, "Mammographic density measured with quantitative computer-aided method: Comparison with radiologists' estimates and BI-RADS categories," *Radiology*, vol. 240, pp. 656–665, 2006.

[32] A. Oliver, J. Freixenet, A. Bosch, D. Raba, and R. Zwiggelaar, "Automatic classification of breast tissue," *Lect. Notes Comput. Sci.*, vol. 3523, pp. 431–438, 2005.

[33] R. Zwiggelaar, I. Muhimmah, and E. R. E. Denton, "Mammographic density classification based on statistical gray-level histogram modelling," in *Proc. Med. Image Understanding Anal. Conf.*, 2005, pp. 183–186.

[34] A. Oliver, J. Freixenet, and R. Zwiggelaar, "Automatic classification of breast density," in *Proc. IEEE Int. Conf. Image Process.*, 2005, vol. 2, pp. 1258–1261.

[35] P. Miller and S. M. Astley, "Classification of breast tissue by texture analysis," *Image Vision Comput.*, vol. 10, pp. 277–282, 1992.

[36] J. W. Byng, N. F. Boyd, E. Fishell, R. A. Jong, and M. J. Yaffe, "Automated analysis of mammographic densities," *Phys. Med. Biol.*, vol. 41, pp. 909–923, 1996.

[37] K. Bovis and S. Singh, "Classification of mammographic breast density using a combined classifier paradigm," in *Proc. Med. Image Understanding Anal. Conf.*, 2002, pp. 177–180.

[38] S. Petroudi and M. Brady, "Breast density segmentation using texture," *Lect. Notes Comput. Sci.*, vol. 4046, pp. 609–615, 2006.

[39] Y. C. Gong, M. Brady, and S. Petroudi, "Texture based mammogram classification and segmentation," *Lect. Notes Comput. Sci.*, vol. 4046, pp. 616–625, 2006.

[40] R. Zwiggelaar, L. Blot, D. Raba, and E. R. E. Denton, "Set-permutation-occurrence matrix based texture segmentation," *Lect. Notes Comput. Sci.*, vol. 2652, pp. 1099–1107, 2003.

[41] R. Zwiggelaar and E. R. E. Denton, "Optimal segmentation of mammographic images," in *Proc. Int. Work. Dig. Mammography*, 2004, pp. 751–757.

[42] H. Li, M. L. Giger, Z. M. Huo, O. I. Olopade, L. Lan, B. L. Weber, and I. Bonta, "Computerized analysis of mammographic parenchymal patterns for assessing breast cancer risk: Effect of ROI size and location," *Med. Phys.*, vol. 31, pp. 549–555, 2004.

[43] H. Li, M. L. Giger, O. I. Olopade, A. Margolis, and M. R. Chinander, "Computerized texture analysis of mammographic parenchymal patterns of digitized mammograms," *Acad. Radiol.*, vol. 12, pp. 863–873, 2005.

[44] L. Blot and R. Zwiggelaar, "Background texture extraction for the classification of mammographic parenchymal patterns," in *Proc. Med. Image Understanding Anal. Conf.*, 2001, pp. 145–148.

[45] D. Raba, A. Oliver, J. Martí, M. Peracaula, and J. Espunya, "Breast segmentation with pectoral muscle suppression on digital mammograms," *Lect. Notes Comput. Sci.*, vol. 3523, pp. 471–478, 2005.

[46] J. C. Bezdek, *Pattern Recognition With Fuzzy Objective Function Algorithms*. New York: Plenum, 1981.

[47] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. New York: Wiley, 2001.

[48] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: A review," *ACM Comput. Surv.*, vol. 31, no. 3, pp. 264–323, 1999.

[49] R. M. Haralick, K. S. Shanmugan, and I. Dunstein, "Textural features for image classification," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-3, no. 6, pp. 610–621, Nov. 1973.

[50] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining Knowl. Discovery*, vol. 2, no. 2, pp. 121–167, 1998.

[51] J. Bi and V. N. Vapnik, "Learning with rigorous support vector machines," *Lect. Notes Comput. Sci.*, vol. 2777, pp. 243–257, 2003.

[52] T. G. Dietterich, "Ensemble methods in machine learning," *Lect. Notes Comput. Sci.*, vol. 1857, pp. 1–15, 2000.

[53] J. J. Rodriguez, L. I. Kuncheva, and C. J. Alonso, "Rotation forest: A new classifier ensemble method," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 10, pp. 1619–1630, Oct. 2006.

[54] R. Collobert, S. Bengio, and Y. Bengio, "A parallel mixture of SVMs for very large scale problems," *Neural Comput.*, vol. 14, pp. 1105–1114, 2002.

[55] X. M. Liu, L. O. Hall, and K. W. Bowyer, "Comments on "A parallel mixture of SVMs for very large scale problems," *Neural Comput.*, vol. 16, pp. 1345–1351, 2004.

[56] J. Kittler, "Feature selection and extraction," in *Handbook of Pattern Recognition and Image Processing*, T. Y. Young and K. S. Fu, Eds. New York: Academic, 1986, pp. 59–83.

[57] J. R. Quinlan, *C4.5: Programs for Machine Learning*. New York: Morgan Kaufmann, 1993.

[58] J. R. Quinlan, "Induction of decision trees," *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, 1986.

[59] J. R. Quinlan, "Bagging, boosting, and C4.5," in *Proc. Nat. Conf. Artif. Intell.*, 1996, pp. 725–730.

[60] J. Cohen, "A coefficient of agreement for nominal scales," *Educ. Psychol. Meas.*, vol. 20, pp. 27–46, 1960.

[61] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, no. 3, pp. 159–174, 1977.

[62] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern Recognit.*, vol. 30, pp. 1145–1159, 1997.

[63] I. Muhimmah, A. Oliver, E. R. E. Denton, J. Pont, E. Pérez, and R. Zwiggelaar, "Comparison between Wolfe, Boyd, BI-RADS and Tabár based mammographic risk assessment," *Lect. Notes Comput. Sci.*, vol. 4046, pp. 407–415, 2006.

[64] S. K. Warfield, K. H. Zou, and W. M. Wells, "Simultaneous truth and performance level estimation (staple): An algorithm for the validation of image segmentation," *IEEE Trans. Med. Imag.*, vol. 23, no. 7, pp. 903–921, Jul. 2004.
[65] D. Kopans, *Breast Imaging*. Philadelphia, PA: Lippincott-Raven, 1998.
[66] C. D. Lehman, E. White, S. Peacock, M. J. Drucker, and N. Urban, "Effect of age and breast density on screening mammograms with false-positive findings," *Amer. J. Roentgenol.*, vol. 173, pp. 1651–1655, 1999.
[67] S. Ciatto, N. Houssami, A. Apruzzese, E. Bassetti, B. Brancato, F. Carozzi, S. Catarzi, M. P. Lamberini, G. Marcelli, R. Pellizzoni, B. Pesce, G. Risso, F. Russo, and A. Scorsolini, "Categorizing breast mammographic density: Intra- and interobserver reproducibility of BI-RADS density categories," *Breast*, vol. 14, no. 4, pp. 269–275, 2005.
[68] F. M. Hall, "Mammographic density categories," *Amer. J. Roentgenol.*, vol. 178, p. 242, 2002.
[69] L. A. Venta and R. E. Hendrick, "Mammographic density categories — Reply," *Amer. J. Roentgenol.*, vol. 178, pp. 242–243, 2002.
[70] A. Oliver, J. Freixenet, R. Martí, and R. Zwiggelaar, "A comparison of breast tissue classification techniques," *Lect. Notes Comput. Sci.*, vol. 4191, pp. 872–879, 2006.

Arnau Oliver received the M.Sc. degree in physics from the Universitat Autònoma de Barcelona, Barcelona, Spain, in 1999, and the Ph.D. degree in information technology from the University of Girona, Girona, Spain, in 2007.

Since 2002, he has been with the Computer Vision and Robotics Group, Department of Electronics, Informatics, and Applications, University of Girona, where he is currently an Assistant Lecturer. His current research interests include pattern recognition and the development of automatic tools for breast cancer detection.

Jordi Freixenet received the M.Sc. degree in computer science from the Polytechnical University of Catalonia, Barcelona, Spain, in 1994, and the Ph.D. degree in computer engineering from the University of Girona, Girona, Spain, in 2000.

He is with the Computer Vision and Robotics Group, University of Girona, Girona, Spain, where he is currently a Lecturer. His current research interests include the field of image processing and computer vision, focusing on medical image analysis, object recognition, image classification, and segmentation. He is also engaged in the improvement in detection and diagnosis of breast cancer.

Robert Martí received the M.Sc. degree in computer science from the University of Girona, Girona, Spain, in 1999, and the Ph.D. degree from the University of East Anglia, Norwich, U.K., in 2002, for his work on image registration applied to multimodal mammography.

He is currently a Lecturer at the University of Girona. His current research interests include medical image analysis, image registration, pattern recognition, and feature extraction techniques specially focusing on mammographic and prostatic data.

Josep Pont received the M.Sc. degree in medicine and the Ph.D. degree in medicine and surgery from the Universitat Autònoma de Barcelona, Barcelona, Spain, in 1980 and 1992, respectively.

He was with the Hospital Universitari de la Vall d'Hebron, Barcelona. He is currently a Radiologist with the Hospital Josep Trueta, Girona, Spain. His current research interests include radiology, specially breast cancer diagnosis using multimodality imaging, using information coming from mammography, echography, and magnetic resonance imaging.

Elsa Pérez received the M.Sc. degree in medicine and surgery from the Hospital Vall d'Hebron, Universitat Autònoma de Barcelona, Barcelona, Spain, in 2000.

She specialized in radiology in the Hospital Bellvitge. She is currently with the Breast Pathology Department, Hospital Josep Trueta, Girona, Spain, where she is engaged in the field of mammography, and ultrasound and magnetic resonance imaging. She has contributed in different studies focusing on breast cancer detection and follow-up as well as in imaging of the reconstructed breast. Her current research interests include the breast tumors vascularization.

Erika R. E. Denton received the M.B.B.S. degree from St. Thomas' Hospital Medical School, London, U.K., in 1989. She trained in radiology at Guys and St. Thomas' Hospital, London, U.K., leading to M.R.C.P. London and F.R.C.R. degrees in 1992 and 1994, respectively.

She then became a Radiology Consultant and a Clinical Lecturer at the King's College Hospital, London, U.K. Since 1999, she has been a Consultant Radiologist with the Norfolk and Norwich University Hospital, Norwich, U.K., where she was the Director of Breast Imaging until 2003. She led the radiology team in Norwich to become one of the three sites developing a new, innovative academy model of education for radiologists. She is also an honorary Senior Lecturer at the University of East Anglia, Norwich, U.K.

Dr. Denton is the Chair of the Royal College of Radiologists Breast Group and the Medical Director of the Picture Archiving and Communications Systems (PACS) Programme at Connecting for Health. She is also the National Clinical Lead for Diagnostic Imaging, Department of Health.

Reyer Zwiggelaar received the Ir. degree in applied physics from the State University Groningen, Groningen, The Netherlands, in 1989, and the Ph.D. degree in electronic and electrical engineering from University College London, London, U.K., in 1993.

He is currently a Senior Lecturer at the University of Wales, Aberystwyth, U.K. He is the author or coauthor of more than 100 conference and journal papers. His current research interests include medical image understanding, especially focusing on mammographic and prostate data, pattern recognition, statistical methods, texture-based segmentation, and feature-detection techniques.