Classification of Mammograms by the Breast Composition

W. R. Silva, D. Menotti
UFOP - Federal University of Ouro Preto
Computing Department
Ouro Preto, MG, Brazil
Email: {welber.ribeiro,menottid}@gmail.com

***Abstract*—Breast cancer produces a high rate of mortality worldwide. Early diagnosis is essential for treatment, however it is difficult to analyse high density breast tissues. Computer-aided diagnosis systems have been proposed to classify the density of mammograms, having as a major challenge to define the features that better represent the images to be classified. In this study, besides comparing them to other techniques, different texture descriptors for the representation of breast tissue density on mammograms are analyzed. In the experiments, 320 mammograms from MIAS Database are used, and the highest accuracy obtained is 77.18% in a 10-fold cross-validation scheme.**

***Index Terms*—mammography, breast cancer, breast density, CAD, texture features**

I. Introduction

Breast cancer is the second leading cause of cancer death in women. Nowadays, it is estimated that one out fourteen women will develop breast cancer during their lifetime. Annually, approximately one million of new cases are diagnosed. An aggravating factor is that $75 - 80\%$ of the patients to be diagnosed are in advanced stages of the disease, which significantly decreases the chances of successful treatment [1].

The diagnosis of this disease is mainly performed using the mammography, a particular form of radiography that uses levels of radiation lower than those of conventional radiography. The mammography records breasts images in order to diagnose the presence of indicative structures of disease [2].

However, the composition of the breast tissue may complicate the detection of lesions. Adipose tissue are less dense and they enable a better detection of lesions. Nonetheless fibroglandular tissue is dense and difficult to detect lesions on it. It is difficult to detect differences between normal tissue and cancerous small dense tissue surrounded by fibroglandular tissue, which makes harder its early diagnosis. Studies such as the ones in [1], [3] show that women with dense breasts have a risk 4-6 times greater of developing breast cancer, and the presence of dense tissue in more than 50% of the breast may be responsible for about one-third of cancer cases.

The reports of mammograms are based on visual analysis of radiologists. In order to standardize them, the American College of Radiology created the standard Bi-RADS [2] that defines breast tissue as:

- BI-RADS I, predominantly fatty, up to 25% of fibroglandular component;
- BI-RADS II, partially fatty, from 26% up to 50% of the volume of breast is fibroglandular tissue;
- BI-RADS III: heterogeneously dense, from 51% up to 75% of fibroglandular tissue;
- BI-RADS IV: extremely dense, more than 75% of fibroglandular tissue.

An example for each kind of breast tissue standard is presented in Figure 1.

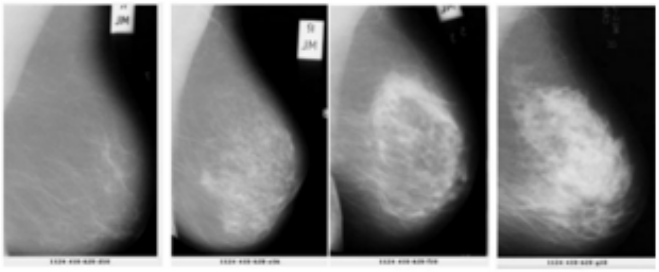

Fig. 1: Standard mammograms of four levels of BI-RADS density. From left to right, BI-RADS I, BI-RADS II, BI-RADS III, and BI-RADS IV, respectively

In order to help radiologists to reduce the variability in the analysis and improve the precision in the interpretation of mammograms, many systems *CAD - Computer-Aided Diagnosis* (Computer Aided Diagnosis) have been proposed [1]. Subashini *et al.* [1] claim that the use of *CADs*, increase the early detection of breast cancer especially in dense breasts, which is also defended by Oliver *et al.* [4]. This last work emphasizes that when facing difficulties for diagnosing lesions in dense breasts, the density classification is important to establish independent strategies for automatic searching for deficiencies in these regions.

Some works approach the use of *Content-based image retrieval* (CBIR) *system* [2], [5] to assist the mammograms diagnosis. Those systems use visual information extracted from the images to retrieve similar image to what is being sought. The stored images are represented/indexed by vectors of features extracted from the images, and for a query image the same vector is obtained and so compared to the vectors of the stored images. The images, limited to a total previously

defined, that have the most similar vectors are returned. In any CBIR approach, the definition of the features set that best represents the density of the breast tissue in mammograms is a challenge.

A system to assist the diagnosis of mammograms should provide:

- Automatic pre-processing;
- Rating of the mammogram according to the breast tissue density;
- Rating of the mammography according to injury;
- Segmentation of the lesion.

In this work, we discuss the classification of mammography according to tissue density by descriptors of texture, comparing them to other techniques, and evaluating the following hypotheses:

- Techniques for representation of breast tissue using principal components analysis are superior to the others;
- The combination of different feature descriptors represents better the breast tissue than when they are individually used;
- The feature extraction of only a portion of mammography is enoguh to classify the tissue;
- Models generated from a small amount of mammography can be sufficiently generalist.

The rest of the paper is structured into six parts. Section 2 presents the approach and the results of related work. In Section 3, we describe the set of mammograms used, and in Section 4, we discuss the particular pre-processing performed in our context. Next, in Section 5, we detail the texture descriptors used, in Section 6 the experiments and, finally, the conclusion and future work are presented in Section 7.

II. Related Work

Several proposals have been made to classify the breast tissue. Here, we highlight the most related to our approach.

Sheshadri & Kandaswamy [6] extract six statistical measures based on the image histogram from 320 mammograms from MIAS Database. That proposed classification approach obtains 80% of accuracy. However they do not report other evaluation measures as the average and standard deviation accuracies of the cross-validation scheme. Without these measures is not possible to know if their model is *overfitted* or not to the training data.

Subashini *et al.* [1] use nine statistics features extracted from the image histograms from *MIAS Database*. Nonetheless they only use 43 mammograms in the experiments and gets an average accuracy of 95.44% using a 3-fold cross-validation scheme.

Oliveira *et al.* [2] apply the technique of principal components analysis in two dimensions on 300×300 pixels *regions of interest* (ROIs), obtaining average accuracies from 83% up to 97% using 10-fold cross-validation scheme. In this study, 5024 mammograms are used from a total of 10,605. Nevertheless only 3,168 mammograms are original. The remaining are scanned copies in different models of scanners. These artificial instances may also bias the ability of generalization of the classifier.

Kinoshita *et al.* [5] extract from 1080 mammograms 88 features related to contour, texture, time, random transform, granulometry and histogram. The resulting feature vector is reduced using the PCA technique and that proposed approach obtains accuracies from 87% up to 93% using a *leave one out* validation scheme.

As we can see from the above examples, a direct and fair comparison between the refenced approaches is impossible. These studies use different databases, different number of instances for training and testing. Besides the pre-processing step and the evaluation measures vary widely at all stages of the approaches. Moreover, the most aggravating factor is that the different databases used do not follow the same standard to classify the tissue, such as BI-RAIDS. Some of them have three classes while others have four. In order to yield a fair and correct comparison of these methods, it is required to implement them using the same methodology and testing using the same databases.

III. Base of mammographic images

The experiments carried out in this work used the image database introduced by [7], which is publicly available for research on http://peipa.essex.ac.uk/info/mias.html. This database contains 322 mammograms of 1024×1024 pixels, which are labeled in accordance with the density of the tissue, instead of four as proposed by the BI-RAIDS, into three classes:

TABLE I: Distribution of instances of the base MIAS

Class	Instances
Fatty	106
Fatty-Glandular	104
Dense-Glanduar	112

IV. Pre-processing

Mammograms, as every acquired information, is highly susceptible to the presence of noise such as the pectoralis muscle, stickers, and any other object not belonging to the breast. Figure 2(a) illustrates such artifats delimited in red which may interfere on the feature extraction process for representing the density of the breast. Therefore a pre-processing step is required to extract a *Region of Interest* (ROI) that contains only the breast tissue. Previous works use different strategies for this filtering process, such as manual [2], semi-automatic [5], or fully automatic [1], [4] segmentation of the ROI. However, none of them assume the existence of breast tissue on the pectoral muscle to be removed, as exemplified in Figure 2(b), and that the removal of that portion of tissue interferes in the mammogram classification where it occurs.

Although it is an essential step in a complete CAD system, the preprocessing is not in the scope of this

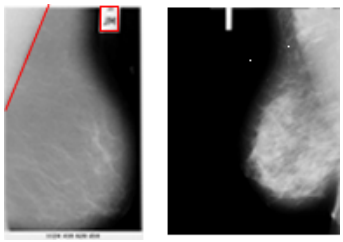

Fig. 2: Pre-processing (a-left) Example of regions to be removed. (b-right) Example of mammography with overlapping of the pectoral muscle.

work. This is the reason why we performed a manual pre-processing of mammograms. the pre-processed images are available in https://github.com/welber/cad_mammography/tree/master/MIAS_PRE_PROC. From now on, please consider that all data we let available in the web are for sake of reproductibiliy of the experiments reported here. Some works such as [2], [5], [6], [4] extract features only from a ROI that represents part of the breast. So, for comparison with these works we extract ROIs of 300 x 300 pixels, and we provide these ROI images in https://github.com/welber/cad_mammography/tree/master/MIAS_300_300.

V. Extraction of features for Classification

As previously stated, the density of breast tissue is a risk factor of the breast cancer developing and the definition of a set of features able to describe the types of breast tissue is a challenging task for the development of a *CAD* system. The visual difference between the tissues in mammography can be defined as the texture of such region. The texture of an image can be represented by statistical descriptors extracted from the histogram of the image intensities or the co-occurrence matrix, besides structural and spectral descriptors.

In order to study and find the best representation of the breast tissue density, we extracted and combined various sets of features:

- Statistics texture descriptors from the histogram of the image intensities [1];
- Statistical texture descriptors from the co-occurrence matrix [8];
- Texture descriptors from the Fourier spectrum [8];
- Invariants Hu moments [8];
- PCA on the matrix image vectored [2];
- 2DPCA [9] on the image array;
- 2DPCA [9] on the co-occurrence matrix [8];

As our experiments show, the combination that generated the best results is statistical texture descriptors applied to the image histogram, also used by [1], combined with statistical texture descriptors from the co-occurrence matrix. The formulation of these descriptors is given below.

A. Statistical texture descriptors from the histogram

As described by [8], let L be the number of possible intensities in an image $M \times N$ pixels, z_i, $i = 0, 1, 2, ...L-1$ their intensity values, and n_i the absolute frequence that z_i occurs in the image. Let also

$$p(z_i) = \frac{n_i}{MN}$$

be the probability of z_i occurs in the image. From these variables, statistical texture descriptors used in this work are defined and extracted as follows:

- **Average** - Average intensity of the image. Regarding mammograms, the more dense tissue is, the higher the average intensity.

$$\mu = \sum_{i=0}^{L-1} z_i p(z_i).$$

- **Standard Deviation** - A measure of contrast intensity grows according to the irregularity of the texture.

$$\sigma = \sqrt{\mu^2}$$

where

$$\mu_2 = \sum_{i=0}^{L-1} (z_i - m)^2 p(z_i).$$

- **Smoothness** - A (relative normalized) measure of smoothness is low to regular intensity and high to irregular.

$$R = 1 \frac{1}{1 + (\frac{\mu^2}{(L-1)^2})}.$$

- **Asymmetry** - Assess whether the intensity levels tend to the dark side or light around the mean.

$$\mu_3 = \sum_{i=0}^{L-1} (z_i - m)^3 p(z_i).$$

- **Uniformity** - It is higher in soft textures, and smaller in more irregular texture.

$$U = \sum_{i=0}^{L-1} p^2(z_i).$$

- **Kurtosis** - Represents how plan is the histogram.

$$\mu_4 = \sum_{i=0}^{L-1} (z_i - m)^4 p(z_i).$$

- **Average histogram** - Estimation of the probability of occurrence of an intensity level [1].

$$AH_g = \frac{1}{L} \sum_{i=0}^{N} (L-1)(i).$$

- **Modified Standard Deviation** - Measure of medium contrast [1]

$$\sigma_m = \sqrt{\sum_{ij} (X_{ij} - \mu)^2 p(X_{ij})}.$$

- **Modified asymmetry** - Assess whether the intensity levels tend to the dark side or light around the mean [1].

$$\sigma_m = \sqrt{\sum_{ij}(X_{ij} - \mu)^2 p(X_{ij})}.$$

B. Statistical descriptors of texture from the co-occurrence matrix

The descriptors based on the histogram do not contain information about the positioning of pixels in relation to other pixels. Thus a mammogram with little dense dots around the breast may be described in the same way as a mammography with a large dense region. In order to preserve the spatial information of the pixel intensity statistics from the co-occurrence matrix of intensity levels can be extracted . For building such matrix, it is needed a Q operator which defines the relative position of the a pixel in relation to the pixel being analyzed. Let $g_{(ij)}$ be an element of a co-occurrence matrix G representing the absolute frequence that the intensities z_i and z_j occur in the position defined by Q [8]. Let also $p_(ij)$ be the probability of $g_{(ij)}$ happens. In other words, $p_(ij)$ can be computed as the ij-th term of G divided by the sum of the elements in G. In the following, we detail the descriptors discussed in [8], which are use in this work.

- **Maximum Probability** - Measure of the largest intensity of G.

$$max(p_{ij}).$$

- **Correlation** - Evaluates how a pixel is related to its neighbor.

$$\sum_{i=1}^{K}\sum_{j=1}^{K}\frac{(m_{i-}r)(j - m_c)p_{ij}}{\sigma_r\sigma_c}.$$

- **Contrast** - Contrast of intensity between a pixel and its neighbor in the image.

$$\sum_{i=1}^{K}\sum_{j=1}^{K}(i-j)^2 p_{ij}.$$

- **Uniformity** - The more uniform the image is, the higher its value.

$$\sum_{i=1}^{K}\sum_{j=1}^{K}p_{ij}^2.$$

- **Homogeneity** - Spatial proximity of the distribution of the G elements.

$$\sum_{i=1}^{K}\sum_{j=1}^{K}\frac{p_{ij}}{1+|ij|}$$

- **Entropy** - Evaluates the randomness of G.

$$\sum_{i=1}^{K}\sum_{j=1}^{K}p_{ij}log_2 p_{ij}$$

VI. Experiments

Once manual pre-processing is performed on 320 mammograms of the mini-MIAS Database (two outliers were removed), features extraction using the different mentioned techniques in the previous section is run. The feature extraction is implemented using Matlab and the source code is available in https://github.com/welber/cad_mammography.

The mammograms used have different labels from the ones of BI-RADS standards and are divided into three classes instead of four. Table II shows the distribution of the instances used in this work.

TABLE II: Distribution of classes

Class	Description	Instances
F	fatty tissue	104
G	glandular tissue	104
D	dense-glandular tissue	112

In order to avoid bias for some features in contrast to others, all features extracted where normalized between -1 and 1, using the following formula,

$$y = \frac{(ymax - ymin) * (x - xmin)}{(xmax - xmin) + ymin},$$

where $ymin$ and $ymax$ stand for the desired range for the new values and $xmin$ and $xmax$ stand for the smallest and largest feature value that is normalized, respectively.

The classification is performed using the Suport Vector Machines algoritm [10], [11] with RBF kernel using the LIBSVM implementation for Matlab [12]. The parameters C and γ were indivually calibrated for different combinations of features tested using a grid search scheme. And the values used for the set of features with the best results are $C = 8,192 = 2^{13}$ and $\gamma = 0.03125 = 2^{-5}$.

The experiments were performed using 10-fold cross-validation scheme. Table III shows the results for some combinations of features. Second and third columns of this table show respectively the mean accuracy obtained from the 10-fold cross-validation scheme using the full instance (Mammography) and the ROIs. As we can observe, the use of ROIs obtained superior results for 5 out 9 feature combinations and showed that its results is similar to ones obtained when using the full mammography.

In order to compare with other studies, we select a small database which is composed of only 44 non-abnormal instances, as suggest in [1]. Table IV presents the results obtained for this experiment where the full mammography is used and the second and third column shown the mean accuracy for 44 and 320 instances, respectively. We can observe that for one case, the small subset of instance obtained significantly high accuracy than the entire subset, and for the other case, the improvement is negligible. So we can see that the result of an classification approach highly depends on the training/testing data. This observation is clue that reinforce

TABLE III: Mean accuracy obtained from the 10-fold cross-validation scheme using 320 instances for full mammography and ROIs

Features	Mamography	ROI
Text. co-occurrence statistics	72.18	70.93
Text. histogram statistics	73.43	71.87
Text. histogram statistics and co-occurence	77.18	73.75
Text. spectral	61.56	71.56
Invariant Hu moments	62.25	69.06
PCA first 5 components	50.62	55.62
PCA first 10 components	50.31	57.81
Text. histogram statistics, co-occurence, and inv. mom.	75.00	74.68
Text. histogram statistics, co-occurence, and spectral	65.31	71.25

what we claimed at the end of our related work section. That is, the comparison of proposed approaches in the literature should be done only using the same evaluation methodology and database.

TABLE IV: Experiments with 44 and 320 instances

Features	44 inst.	320 inst.
Text. histogram statistics	74.41	73.43
Text. statistics histogram and co-occurrence	88.37	77.18

TABLE V: Experiments with techniques of principal component analysis (PCA)

# PCs	PCA	2DPCA	2DPCA & Text.
First 5	55.62	59.68	72.18
First 6	56.56	59.68	72.50
First 7	57.18	59.37	72.81
First 8	58.75	59.68	72.50
First 9	60.62	58.75	72.50
First 10	57.81	58.43	72.18
First 11	57.50	58.75	72.81
First 20	56.25	55.93	73.43

Table V presents the results of experiments using the PCA technique. In the second, third and fourth columns of this table we can see by varying the number of principal components (PCs) the accuracies obtained by approaches using PCA, 2DPCA, and 2DPCA combined with the statistical features of texture. We can also observe that in any of these experiments that the use of statistical texture descriptors significantly improved the accuracies. Moreover, the employ of the PCA technique does not bring to us benifits regarding the accuracy obtained.

In all experiments, the combination of features from differente natures obtained greater accuracies than when they were used alone. And the features that best represented the density of breast tissue were the statistical descriptors of texture of image histogram intensity and co-occurrence matrix, obtaining the mean accuracy of 77.18%. Table VI shows the confusion matrix for this set of features.

TABLE VI: Confusion Matrix

-	D	F	G
D	97	0	10
F	3	91	8
G	12	13	86

VII. Conclusion and future work

In this work, we evaluated indivually and combining different sets of features for breast tissue classification. The highest mean accuracy, in a 10-fold cross-validation scheme, obtained was 77.18% that used the combination of statistical features extracted from the histogram and co-occurrence matrix. Moreover, from the experiments, we can conclude that: 1) Techniques for representation of breast tissue using principal components analysis obtain worse accuracies; 2) The feature extraction of only a portion of mammography is enough to classify the tissue; 3) Models generated using a database and specific evaluated methodology cannot be compared in other context.

As demonstrated by the confusion matrix of all the instances incorrectly classified, only three of them did not belong to the class of intermediate density or were incorrectly classified as belong to that class. Features that are needed to define better the differences that class to the other. New features are required to better discriminates of intermediate density class to the other classes. Future studies aim to analyse whether on grounds that are labeled into four classes according to BI-RAIDs this error is minimized.

Also future work we plan to develop more elaborated classification methods combining other texture descriptors using committee of classifiers. Other factors that can improve the results are the use of features that take into account the different sizes of breast cancer, and to carry out a pre-processing that consider breast tissue over the pectoralis muscle before of the removal step.

VIII. Acknowledgements

The authors would like to thank FAPEMIG, CAPES and CNPq for the financial support.

REFERENCES

[1] T. Subashini, V. Ramalingam, and S. Palanivel, "Automated assessment of breast tissue density in digital mammograms," *Computer Vision and Image Understanding*, vol. 114, no. 1, pp. 33–43, 2010.

[2] J. E. E. D. Oliveira, A. M. C. Machado, G. C. Chavez, A. P. B. Lopes, T. M. Deserno, and A. de A. Araújo, "Mammosys: A content-based image retrieval system using breast density patterns," *Computer Methods and Programs in Biomedicine*, vol. 99, no. 03, pp. 289–297, 2010.

[3] N. F Boyd, H. Guo, L. J Martin, L. Sun, J. Stone, E. Fishell, R. A Jong, G. Hislop, A. Chiarelli, S. Minkin, and M. J Yaffe, "Mammographic density and the risk and detection of breast cancer," *The New England Journal of Medicine*, vol. 356, no. 3, pp. 227–236, 2007.

[4] A. Oliver, X. Llad, E. Prez, J. Pont, E. R. E. Denton, J. Freixenet, and J. Mart, "A statistical approach for breast density segmentation," *Journal of Digital Imaging*, vol. 23, no. 5, pp. 527–537, 2010.

[5] S. K. Kinoshita, P. M. de Azevedo-Marques, R. R. Pereira, J. A. H. Rodrigues, and R. M. Rangayyan, "Content-based retrieval of mammograms using visual features related to breast density patterns," *Journal of Digital Imaging*, vol. 20, no. 2, pp. 172–190, 2007.

[6] H. S. Sheshadri and A. Kandaswamy, "Experimental investigation on breast tissue classification based on statistical feature extraction of mammograms," *Computerized Medical Imaging and Graphics*, vol. 31, no. 1, pp. 46–48, 2007.

[7] J. Suckling, "The mammographic image analysis society digital mammogram database exerpta medica," *International Congress Series*, no. 1069, pp. 375–378, 1994.

[8] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, 3rd ed. Prentice Hall, 2008.

[9] J. Yang, D. Zhang, A. F. Frangi, and J. yu Yang, "Two-dimensional pca: A new approach to appearance-based face representation and recognition," *IEEE Transactions on Pattern Analysis Machine and Intelligence*, vol. 26, no. 1, pp. 131–137, 2004.

[10] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, pp. 273–297, 1995.

[11] S. Theodoridis and K. Koutrambas, *Pattern Recognition*, 4th ed. Academic Press, 2009.

[12] C.-C. Chang and C.-J. Lin, "Libsvm: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 1–27, 2011.