

Breast Lesions Classification applied to a reference database

Júlia E. E. de Oliveira ^{*}, Thomas M. Deserno ^{**} and Arnaldo de A. Araújo ^{*}

^{*} *Department of Computer Science, Federal University of Minas Gerais, Belo Horizonte, Brazil*

julia@dcc.ufmg.br
arnaldo@dcc.ufmg.br

^{**} *Department of Medical Informatics, Aachen University of Technology, Aachen, Germany*

deserno@ieee.org

Abstract: In this paper, we propose a method to classify two breast lesions – mass and calcification, despite their benignancy or malignancy. Applied to a reference database that provides images with the ground truth set, the evaluation of the method is easy and trustful. This database, from the IRMA project, was developed from the union from of: The Digital Database for Screening Mammography (DDSM), The Mammographic Image Analysis Society Digital Mammogram Database (MIAS), the Lawrence Livermore National Laboratory (LLNL), and routine images from the Rheinische-Westfälische Technische Hochschule (RWTH) Aachen. Wavelet transforms, like Haar and Daubechies, in a one and two-level decomposition, were used to characterize the normal regions and the regions containing the lesion – mass and calcification. The detail coefficients obtained with this wavelet decomposition were used to train and test a support vector machine classifier with linear kernel. A result of 89.6% of accuracy and 100% of specificity points this method as promising, allowing additional studies for its improvement.

Key words: : breast lesion classification, IRMA database, support vector machine, wavelet transform.

INTRODUCTION

Breast cancer represents one of the main causes of death among women in occidental countries (Brazil National Cancer Institute, <http://www.inca.gov.br>). Mammography is the best method of screening for breast cancer because it can show lesions in their initial phases. Masses and calcifications are examples of lesions present in mammographies, as can be seen in Figure 1, and their diagnostic is made by a radiologist.

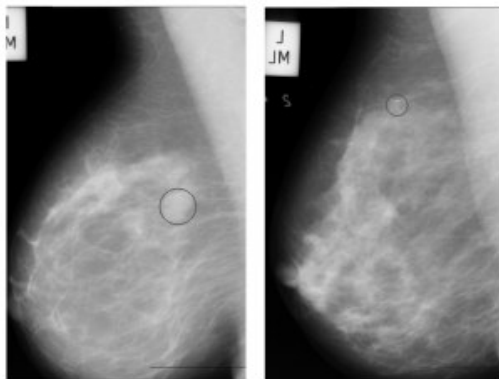


Figure 1. *On the left, mammography containing mass. On the right, one containing calcification.*

Due to the continuous advances of technology there is an increase in the interest of computer-aided diagnosis (CAD) systems [Leh 05] [Doi 07] [Ran 07]. The computer is used as a second opinion such that the performance by computers is complementary to that by radiologists. An effective CAD system, i.e., a system that clearly identifies position, size and staging of lesions like masses and calcifications in x-ray mammographies, must be evaluated using a large number of reference images with approved diagnostics (ground truth). Besides, such a system should provide the analysis and classification of these lesions.

A brief revision of the literature shows the techniques that have been used in attempts to classify breast lesions. For instance, Eltonsy et al. [Elt 07] presented a technique for automatic mass detection in mammographies. Images were cropped to regions containing malignant and benign masses and normal regions. Morphological characteristics were extracted from these regions of interest (ROI) and a minimum distance classifier was used in 540 images containing malignant masses, 270 images containing benign masses and 164 normal images from the DDSM database (The Digital Database for Screening Mammography) [Hea 98]. The performance achieved with the proposed CAD scheme was 92.1% sensitivity

for malignant masses at 5.4 false-positive per image.

Verma [Ver 08] proposed a novel algorithm for the classification of mass abnormalities in digitized mammograms. 100 benign and 100 malignant cases from the DDSM database were used for training and testing the proposed classifier. All the images were cropped to discard black areas and patient information. For the composition of features vectors, grey level-based features were extracted, like average histogram, average grey, contrast, energy, modified energy, entropy, modified entropy, standard deviation, modified standard deviation, skew and modified skew. The novel algorithm proposed for classification of all these features is based on the introduction of additional neurons in the hidden layer for benign and malignant classes and a weight adjustment technique for the calculation of weights. Results show 94% classification accuracy on test set and 100% classification accuracy on training set.

Zwiggelaar et al. [Zwi 04] aimed the detection of linear structures and their classification into vessels, spicules, ducts, etc., in a way that this provided anatomical information can be used to improve the specificity of automatic abnormality detection. From MIAS database (The Mammographic Image Analysis Society digital mammogram database) [Scu 96], 15 mammographies were used and from them, 274 linear structures were cropped. Synthetic data was used as a comparison to the performance of four methods: line operator, orientated bins, Gaussian derivatives and ridge detection. The classification was done using a Gaussian model and all the methods were evaluated by receiver operating characteristics (ROC) [Faw 04] curves.

The main efforts for breast lesions classification, as can be seen in the presented works, are the differentiation between benign and malignant cases and the classification of calcifications and masses separately.

In a way to overcome the shortcomings of the existing works, we propose a method for classification of two breast lesions: calcification and mass, using wavelets for breast lesions characterization and support vector machine (SVM) for the task of classification. The evaluation is based on the IRMA (Image Retrieval in Medical Applications) database [Leh 05] [Des 07].

The remainder of this paper is broken into five sections. Section 1 presents the database used for the evaluation of our algorithm. Section 2 introduces how breast lesions were characterized through the Haar and the Daubechies wavelet transform. In Section 3, we present a brief description of the principles of the support-machine classifier. In Section 4, we present the experiments and results, and in Section 5, we discuss the results and state the conclusion of the work.

1. IRMA Database

The IRMA project (<http://www.irma-project.org>) aims at developing and implementing high-level methods for content-based image retrieval (CBIR) systems with prototypal applications to medico-diagnostic tasks on radiological image archive [Leh 04]. There are currently more than 30,000 diagnostic images with available ground truth information in the IRMA database. They are used for image retrieval and computer-aided diagnosis [Leh 05] [Des 07].

Regarding mammography, there are more than 10,000 images in the database [Oli 07][Oli 08]. This mammography database was developed from the union of: The Mammographic Image Analysis Society Digital Mammogram Database (MIAS), The Digital Database for Screening Mammography (DDSM), the Lawrence Livermore National Laboratory (LLNL), and routine images from the Rheinische-Westfälische Technische Hochschule (RWTH) Aachen.

In IRMA, all the images are coded according to a mono-hierarchical, multi-axial coding scheme [Leh 03]. The IRMA code for mammographies is represented by four axes that describe:

- technique: image modality, i.e., x-ray plain radiography;
- direction: body orientation, i.e., cranio-caudal or medio-lateral;
- anatomy: body region examined, i.e., right or left breast;
- biosystem: biological system examined, i.e., tissue density, tumor staging, and lesion description, as can be seen in Tables 1, 2, and 3.

IRMA code	Tissue density description
d	fat transparent system
e	fibroid glands system
f	heterogeneously dense system
g	extremely dense system

Table 1. IRMA code for breast tissue density.

IRMA code	Tumor staging description
0	unspecified
1	normal
2	benign
3	probably benign
4	suspiciously abnormal
5	malignant

Table 2. IRMA code for tumor staging.

This codification sets the ground truth of the images, allowing an accurate evaluation of our proposed algorithm.

In addition, the IRMA framework provides a database that hosts images as well their derived features. This is used to store one or more lesions descriptions, in form of:

- circle: described by its center coordinates and radius;
- contour points: a list of (x,y)-coordinates;

IRMA code	Type of lesion description
0	unspecified
1	calcification, unspecified
2	microcalcification
3	macrocalcification
4	circumscribed mass
5	spiculated mass
6	other mass
7	architectural distortion
8	asymmetry

Table 3. IRMA code for type of lesion.

- chain code: a starting coordinate (x,y) followed by a sequence of numbers describing the direction to the adjacent contour point;
- masking image: a binary image with the same (x,y) dimensions, as the mammography, where 0 and 1 denote “background” and “lesion”, respectively.

As can be seen in Figure 2, this visual identification of the lesion along with the IRMA code aid the extraction of the ROI containing the lesion besides, as already said, setting the ground truth of all the images.

2. Wavelets for breast lesions characterization

The representation of masses and calcifications can be made by a vector of characteristics, also referred to as numerical signature. Wavelets are good for the extraction of the vector of characteristics because they can represent images in a smaller scale with a minimum amount of values, besides they hardly depend on the resolution of the original image [Dau 90][Uns 96].

In the literature, wavelet transform was used by Hamad et al. [Ham 06] to investigate the better type of wavelet transform – Daubechies, Symlet, Coiflet, and biorthogonal – and its optimal level of decomposition

that detect microcalcifications in breast cancer. The approximation coefficient was discarded for the reconstruction of the mammographies in a way to highlight the microcalcifications. From the DDSM database, they used 10 mammographies containing microcalcifications. As results they point that the use of wavelets which have functions similar to the profile shape of microcalcifications improve the detection.



Figure 2. Mammography of IRMA biosystem code “d24” - dense breast tissue containing a benign circumscribed mass identified by a circle.

Sentelle et al. [Sen 02] proposed the detection and enhancement of digital mammograms so that microcalcifications may be brought to attention of a radiologist. 20 mammographies containing malignant calcifications and 5 normal mammographies, from DDSM database, were preprocessed to discard pectoral muscle and areas that are not tissue. Afterward, in a four-level decomposition they applied the biorthogonal wavelet transform and the lowest approximation coefficients were removed, so when the reverse wavelet transform was performed for the reconstruction of the image, the microcalcifications were highlighted. In dense tissues, 80% detection rate with less than four false positives per image was obtained.

As wavelet transform was used for detection and segmentation of masses and calcifications, separately, we evaluate the use of the Haar and the Daubechies wavelet transform for breast lesions – mass and calcification – characterization.

Wavelet transform can well represent images in a smaller scale with a minimum amount of values, besides it hardly depends on the resolution of the original image [Dau 90][Uns 96]. Wavelets are mathematical functions that process data in different scales and resolutions, representing them in a coarse way together with their details [Gra 95][Add 02]. An image decomposition using wavelet functions is

known as discrete wavelet transform and this kind of transform reveals not only the frequency attributes of the images but also the spatial ones.

The Haar and the Daubechies wavelet transform can represent the difference between masses and calcifications through the gray levels. This type of analysis using wavelet transform consists in the decomposition of the original image in approximation and detail coefficients. Approximation represents the low frequency components of the images and the details represent the high frequency components (Fig. 3).

Figure 4 presents a two-level decomposition through the Haar wavelet transform. The approximation coefficient imparts nuance and the detail coefficients – horizontal, vertical, and diagonal – give the image its identity.

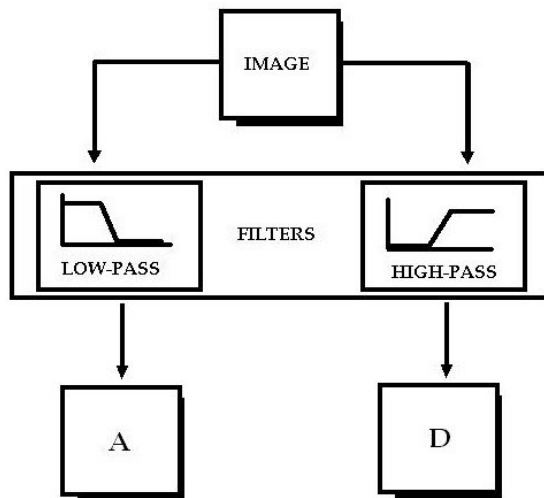


Figure 3. Discrete wavelet decomposition.

3. Support vector machine for classification

Classification identifies categories that can represent masses and calcifications. The task of the classifier is to use the vector of characteristics to assign the ROI to a category – lesion or normal – determining the probability for each of the possible categories.

When the category of the image is already set, the classification is called supervised learning [Dud 01]. That means an extern agent is used to indicate the desired answers for the pattern input. The classifier is then trained using a large set of labeled training samples.

The portioning of the data into subsets for training and testing is called cross-validation [Koh 95], and the type called holdout validation chooses the observations randomly from initial sample to form the validation data (test data), and the remaining observations are retained as the training data.

The use of support vector machine (SVM) as classifier may be found in the literature in the last years [Naq 04][Dat 05][Li 06].

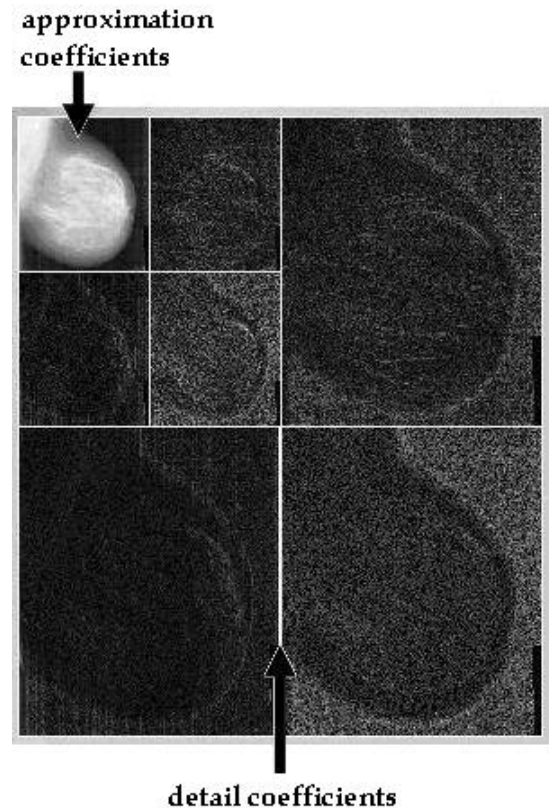


Figure 4. Example of a two-level decomposition through the Haar wavelet transform.

For instance, Arodz et al. [Aro 05] in a method evaluation of recognition of suspicious anomalies in mammographies, used SVM with linear, quadratic and polynomial kernels as classifier. Even though it did not perform as well as the other used classifier, the AdaBoost [Fre 99], the authors pointed out that a better characterization of the breast lesions should be the solution for this problem.

In [Cam 04], Campanini et al. developed a technique for mass detection in mammographies for the use in CAD systems and applied it in 1,400 mammographies (800 containing masses and 600 normal) from DDSM database. For feature extraction, the Haar wavelet was used in the cropped images and two stages of classification, both using SVM, were applied: the first stage for classification is based on distance and the second one intends the reduction of false candidates. Results evaluated by the sensitivity of the method points almost 80% of masses correctly identified with a false-positive rate of 1.1 marks per image.

For our study, we have chosen the use of SVM for the task of classification because they have [Bur 98]:

- good capacity of generalization: efficient classification of data that does not belong to the set used for training;

- great robustness in higher dimensions: they work well with images;
- a theory well defined: support vectors are a well established theory base in mathematics and statistics.

The purpose of support vectors is to find the most adequate hyperplane that is able to separate the groups in a way that the cases belonging to a certain category stay at one side of the plane and the other cases stay at the other side. Support-vectors are trained using linear and non-linear kernels, like Gaussian, polynomial and quadratic [Hsu 02] (Fig. 5).

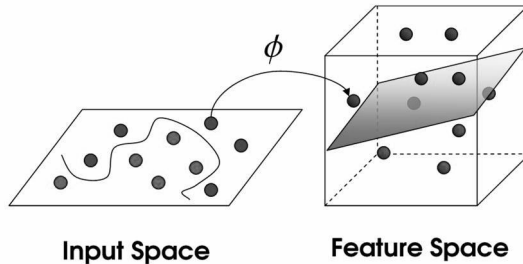


Figure 5. Example of SVM classification.

4. Experiments and Results

A diagram identifying the major components of the proposed method is given in Figure 6.

We evaluated 96 mammographies from IRMA database: 48 mammographies containing lesions – mass and calcification – and 48 normal mammographies.

As all the images in the IRMA database have the identification of the lesions by a circle, this helped us to manually select only the ROI containing mass or calcification, and also to select normal regions of the same size of the ones containing a lesion.

All algorithms were implemented using MatLab (Matrix Laboratory, <http://www.mathworks.com>), using the Wavelet and Bioinformatics toolboxes. After the ROI selection, the steps for this implementation were:

- portioning the set of ROIs, using the holdout cross validation;
- extraction of the horizontal, vertical and diagonal detail coefficients of the Haar and the Daubechies wavelet transform for each ROI, in a one and two level decomposition;
- training and classification of each of detail coefficient, separately, using SVM with linear kernel;
- evaluation through a statistical method.

Based on Everitt [Eve 94], Petroudi et al. [Pet 03], and Zrimec et al. [Zri 07], for an analysis of the proposed method, the sensitivity, specificity and accuracy tests were applied. In our case, sensitivity determines the proportion of ROIs containing lesions

that have been detected as containing lesion – mass or calcification. Specificity indicates the proportion of normal ROIs that have been detected as normal. The accuracy of this classifier is the percentage of correctly classified ROIs over the ground truth of total ROIs in that category (normal or lesion).

The best results were the ones using the vertical detail coefficient of both Haar and Daubechies wavelet transform, in a one and two level decomposition, as can be seen in Table 4.

	sensitivity	specificity	accuracy
Haar one level	79.2%	87.5%	83.3%
Haar two level	83.3%	95.8%	89.6%
Daubechies two level	70.8%	100%	85.4%

Table 4. Results for the proposed method using the vertical detail coefficient.

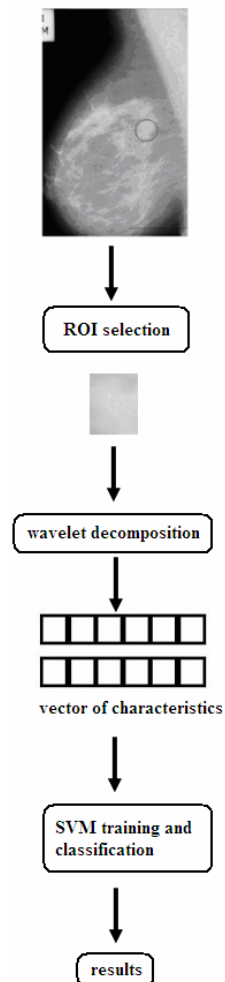


Figure 6. Experimental method overview.

5. Discussion and Conclusion

This paper has evaluated a method of classification of two breast lesions – mass and calcification, without differentiating them between benign or malignant. The database used, from the IRMA project, provides mammographies with the ground truth already set, as all the images were previously checked by an expert radiologist. In addition, all the images that contain a certain lesion had its identification by a circle.

For the evaluation of this method, we chose to use together the Haar and the Daubechies wavelet transform in different levels of decomposition for the breast lesions characterization, and support vector machine for classification, as they already got good results for breast lesions detection and classification [Wan 98][Sen 02][Aro 05].

Our study demonstrated that although all the wavelet detail coefficients – horizontal, vertical and diagonal - give the identity of the images, the vertical detail coefficient was the one that really characterized the ROI containing lesions and the normal ones. This may have happened because even though the detail coefficients represent the high frequency components, their composition is mathematically different [Dau 90], allowing the vertical detail coefficient in a one and two-level decomposition to capture in details the gray level difference between ROIs containing lesion and normal ROIs.

Results showed 83.3%, 89.6%, and 85.4% of accuracy using the Haar wavelet transform in a one level decomposition, the Haar wavelet transform in a two level decomposition, and the Daubechies wavelet transform in a two level decomposition, respectively, with the linear kernel of the SVM classifier. A 100% of correct normal regions classification was obtained with the use of the Daubechies wavelet transform in a two level decomposition. This denotes that the vertical detail coefficient was able to well characterize the regions in a way to let the support vectors to set a separate cluster of the lesions regions versus the normal regions.

For the improvement of the studied method, we shall apply it to a larger set of data, trying other methods of cross-validation, like the k-fold and leave-one-out [Dud 01] types.

Future works include automatic selection of the ROIs containing the lesion, and as the purpose of the proposed method was not the differentiation between malignant of benign lesions, we should try to do this differentiation applying shape attributes to capture all the extension of the lesion.

All things considered, we conclude stating that the evaluation of the method of this study applied to the IRMA database is promising and further studies are needed to integrate our method in a computer-aided system of mammographies.

ACKNOWLEDGMENT

This research is supported by CAPES and CNPq, Brazilian research funding agencies. The IRMA project is funded by the German Research Foundation (DFG), Le 1108/4, and Le 1108/9.

REFERENCES

- [Add 02] Addison P.S. (2002). The illustrated wavelet transform handbook. USA: Taylor and Francis Group.
- [Aro 05] Arodz T., Kurdziel M., Sevre E.O.D., Yuen D.A. (2005). Pattern recognition techniques for automatic detection of suspicious-looking anomalies in mammograms. *Computer Methods and Programs in Biomedicine*, 79:135-149.
- [Bur 98] Burges C. (1998). A tutorial on support-vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2:121-167.
- [Cam 04] Campanini R., Dongiovanni D., Iampieri E., Lanconelli N., Masotti M., Palermo G., Ricarddi A., Roffilli M. (2004). A novel featureless approach to mass detection in digital mammograms based on support-vector machines. *Physics in Medicine and Biology*, 49:961-975.
- [Dat 05] Datta R., Li J., Wang J.Z. (2005). Content-based image retrieval: approaches and trends of the new age. *MIR'05 Proceedings of the 7th ACM SIGMM International Workshop on Multimedia Information Retrieval*, 253-262.
- [Dau 90] Daubechies I. (1990). The wavelet transform, time-frequency localization and signal analysis. *IEEE Transactions on Information Theory*, 36:961-1005.
- [Des 07] Deselaers T., Müller H., Clough P., Ney H., Lehmann T. (2007). The CLEF 2005 automatic medical image annotation task. *International Journal of Computer Vision*, 74(1):51-58.
- [Doi 07] Doi K. (2007). Computer-aided diagnosis in medical imaging: historical review, current status and future potential. *Computerized Medical Imaging and Graphics*, 31, 198-211.
- [Dud 01] Dudda R.O., Hart P.E., Stork D.G. (2001). Pattern classification. Canada: John Wiley Sons.
- [Elt 07] Eltonsy N.H., Tourassi G.D., Elmaghraby A.S. (2007). A concentric morphology model for the detection of masses in mammography. *IEEE Transactions on Medical Imaging*, 26(6):880-889.
- [Eve 94] Everitt B. (1994). Statistical methods in medical investigations. Halsted Press, London.
- [Faw 04] Fawcett T. (2004). *ROC Graphs: notes and practical considerations for researches*. Kluwer Academic Publishers, 1-38.
- [Fre 99] Freund Y., Schapire R.E. (1999). A short introduction to boosting. *Journal of Japanese Society for Artificial Intelligence*, 14(5):771-780.
- [Gra 95] Graps A. (1995). An introduction to wavelets. *IEEE Computational Science and Engineering*, 2(2):50-61.
- [Ham 06] Hamad N.B., Taouil K. (2006). Exploring

wavelets subband decomposition toward a computer-aided detection of microcalcification in breast cancer. *The 2nd International Conference on Distributed Frameworks for Multimedia Applications* 2006; 1-8.

[Hea 98] Heath M., Bowyer K.W., Kopans D. et al. (1998). *Current status of the digital database for screening mammography*. In: *Digital Mammography*, Kluwer Academic Publishers, 457-460.

[Hsu 02] Hsu C.W., Lin C.J. (2002). A comparison of methods for support vector machine. *IEEE Transactions on Neural Networks*, 13(2):415-425.

[Koh 95] Kohavi R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *IJCAI*, 1137-1145.

[Leh 03] Lehmann T., Schubert H., Keysers D., Kohnen M., Wein B. (2003). The IRMA code for unique classification of medical images. *Proceedings SPIE*, 5033:440-451.

[Leh 04] Lehmann T.M., Gueld M.O., Thies C., Fischer B., Spitzer K., Keysers D., Ney H., Kohnen M., Schubert H., Wein B. (2004). Content-based image retrieval in medical applications. *Methods of Information in Medicine*, 43(4):354-361.

[Leh 05] Lehmann T., Gueld M., Deselaers T., Keysers D., Schubert H., Spitzer K., Ney H., Wein B. (2005). Automatic categorization of medical images for content-based image retrieval and data-mining. *Computerized Medical Imaging and Graphics*, 29(2): 143-155.

[Li 06] Li J., Allinson N., Tao D., Li X. (2006). Multitraining support vector machine for image retrieval. *IEEE Transactions on Image Processing*, 15(11):3597-3601.

[Naq 04] Naqa I.El, Yang Y., Galatsanos N.P. (2004). A similarity learning approach to content-based image retrieval: application to digital mammography. *IEEE Transactions on Medical Imaging*, 23(10):1233-1244.

[Oli 07] Oliveira J.E.E., Gueld M., Araújo A.A., Ott B., Deserno T. (2007). Building a standard reference database for computer-aided mammography diagnosis. *3rd International Conference of the Brazilian Association for Bioinformatics and Computational Biology (X-Meeting)*.

[Oli 08] Oliveira J.E.E., Gueld M., Araújo A.A., Ott B., Deserno T. (2008). Towards a reference database for computer-aided mammography. *Proceedings SPIE Medical Imaging*, 6915:69151Y.

[Pet 03] Petroudi S., Kadir T., Brady M. (2003). Automatic classification of mammographic parenchymal patterns: a statistical approach. *Proceedings of the 25th International Conference of the IEEE EMBS*.

[Ran 07] Rangayyan R.M., Ayres F.J., Desautels J.E.L. (2007). A review of computer-aided diagnosis of breast cancer: toward the detection of subtle signs. *Journal of the Franklin Institute*, 344: 312-348.

[Scu 96] Suckling J. et al. (1996). The Mammographic image analysis society digital mammogram database. *Exerpta Medical International Congress Series*, 1069:375-378.

[Sen 02] Sentelle S., Sentelle C., Sutton M.A. (2002). Multiresolution-based segmentation of calcifications for the early detection of breast cancer. *Real Time Imaging*; 8:237-252.

[Uns 96] Unser M. (1996). A review of wavelets in biomedical applications. *Proceedings of the IEEE*, 84:626-638.

[Ver 08] Verma B. (2008). Novel network architecture and learning algorithm for the classification of mass abnormalities in digitized mammograms. *Artificial Intelligence in Medicine*, 42:67-79.

[Wan 98] Wang T.C., Karayiannis N.B. (1998). Detection of microcalcifications in digital mammograms using wavelets. *IEEE Transactions on Medical Imaging*, 17(4):498-509.

[Zri 07] Zrimec T., Wong J.S. (2007). Improving computer aided disease detection using knowledge of disease appearance. *Studies in health technology and informatics*, 129: 1324-1328.

[Zwi 04] Zwigglaar R., Astley S.M., Boggis C.R.M., Taylor C.J. (2004). Linear structures in mammographic images: detection and classification. *IEEE Transactions on Medical Imaging*, 23:1077-1086.