Classification of Breast Tumors Through Image Mining Techniques

Lizianne P. Marques Souto[1,2], Thiago K. L. dos Santos[1], Marcelino Pereira S. Silva[1]

[1]Universidade do Estado do Rio Grande do Norte (UERN)

[2]Universidade Federal Rural do Semiárido (UFERSA)

{lizianne.priscilla, thiagokleyton}@gmail.com

marcelinopereira@uern.br

Abstract. *In this paper a methodology and a computer-aided diagnosis system for detection of breast cancer are proposed. The approach involves Image Processing resources to extract morphological features from tumors in mammograms and Image Mining to classify them as benign or malignant. Images from BCDR repository were used for the experiments. The results showed the efficacy of the proposed method and system, which reduced the false positive and false negative rates, and allowed a more efficient decision-making process.*

1. Introduction

The increase in the occurrence of new cases of breast cancer has been considered one of the serious public health problems in the world. Developing countries are the most affected by this disease followed by a high mortality rate due to the detection at advanced stages. The way to combat the breast cancer is to do regularly the breast self-examination, the clinical exam by a doctor and imaging exams, such as mammography. The early detection, through treatments, increase the chances of patient survival. Among all screening methods currently available, mammography is the most reliable to detect tumors at the initial stage, sized from 1mm, while the breast self-exam can only detect tumors from 1.5cm [Leite et al. 2011].

The mammogram analysis is generally performed by radiologists. This task demands specific training and a lot of experience of the professional, due to factors such as low image quality, size and morphological variation of the lesions that can not provide a precise and uniform evaluation. For such reasons, about 10% to 30% of breast lesions are not identified on mammograms [Calas et al. 2012]. It is also estimated that the sensitivity, i.e. detection of true-positive cases, of radiologists in breast cancer screening is between 65% and 75% [Skaane et al. 1997].

A possible solution is the use of computer-aided diagnosis (CAD) systems that provides a double reading of images, increasing up to 15% the sensitivity in the detection of breast cancer [Thurfjell et al. 1994]. CADs decrease the uncertainty of the specialist in the diagnosis, providing a second opinion about the case. These systems are useful to avoid distortions in the interpretation of lesions, incorrect treatments and unnecessary surgical biopsies.

In general, CAD systems improve the image quality, and consequently the visualization and localization of suspicious lesions, extract features from images and classify

the mammographic findings according to their probability of malignancy. In this paper the development of a methodology and a CAD system to assist radiologists in diagnosis of breast cancer is proposed. The methodology uses techniques of Image Processing and methods of Data Mining that allow distinguishing the tumors as benign or malignant, according to their morphological features.

2. Computer-Aided Diagnosis: Related Work

Several CAD methodologies for diagnosis of breast cancer are found in the literature. An example is the work in [Asad et al. 2011] that had as objective the classification of breast tumors in malignant and benign categories. A set of 33 mammograms from Mammographic Image Analysis Society (MIAS) database were segmented by Local Thresholding to separate the ROI (region of interest) from images. Seven geometric attributes were extracted and then classified using Kohonen Neural Networks.

In another approach, mammograms were classified using geometric attributes and features extracted from the edge of regions for the detection of breast cancer [Surendiran and Vadivel 2012]. From 940 images of DDSM database, 17 attributes were extracted. To identify and classify regions of interest found in mammograms, the Threshold method and CART algorithm were used. In [Radovic et al. 2013] the pectoral muscle was removed from the image, then the region of interest was identified in the preprocessing phase using Local Thresholding. Twenty texture attributes were extracted from 322 images of the MIAS database and used to classify them as normal (no tumor) or abnormal (with tumor). Support Vector Machine (SVM), Naive Bayes, K-Nearest Neighbor (KNN), Logistic Regression, Decision Tree (C4.5), Random Forest and Multilayer Perceptron were used in the classification.

The work in [Liu and Tang 2014] focused on methods of feature selection for the classification of breast tumors. From 826 images of DDSM repository, 31 features of shape and texture were extracted. The authors developed the Spatial Fuzzy C-Means Clustering method that provides the best initialization for segmentation of the ROI. This method was based on the Level Set, which is originated from the Active Contour Model (Snakes) for identifying the edges of the regions. Several methods of feature selection were investigated, and the most efficient selected 12 features, which were classified by Support Vector Machine (SVM) and KNN.

3. Breast Tumor Morphology

The disorderly growth and multiplication of cells result in tumor formation. When malignant, the tumor is considered cancer, tends to invade adjacent tissues and therefore has an irregular shape and indistinct edge [Alvarenga et al. 2003]. On the other hand, benign tumors tend to have rounded or oval shape and circumscribed edge. Thus, the tumors usually are visually distinguishable according to the parameters:

- Size: small tumors can not be detected through clinical exams, and can only be identified through mammograms;
- Shape: tumors may have round, oval or irregular shape;
- Margin: malignant tumors generally have an irregular or undefined contour;
- Density: it is possible to classify the tumor density in relation to the surrounding normal glandular tissue. Malignant tumors tend to have high density and appear in the image as white areas.

4. Image Processing

In Image Processing, techniques are applied to improve the image quality, highlighting the edges of objects and eliminating noise acquired in image acquisition [dos Santos Romualdo et al. 2009]. This process can includes operations such as image preprocessing, identification of regions of interest, feature extraction and classification (recognition) of objects.

4.1. Preprocessing

In preprocessing phase, techniques are applied to improve image visualization and interpretation by the observer. There are many enhancement techniques, which choice and use are directly related to the context and to the analyst preference in respect to a "good image" [Gonzalez and Woods 2008].

In this work the Negative Transformation technique was used. This technique reverses the image histogram and, consequently, the grayscale of the image. It highlights white or gray details in dark regions of an image [Gonzalez and Woods 2008].

4.2. Image Segmentation

Segmentation is the subdivision of an image into its constituent regions until the objects of interest are identified [Gonzalez and Woods 2008]. This technique is very important in pattern recognition tasks [Cuadros et al. 2012].

Among several segmentation techniques, the region growing (RG) was used in this work due to its wide application in the literature. This method is based on the fact that an image is composed by a set of regions and them, in turn, are formed by a set of pixels. The growth results in grouped regions according to predetermined parameters such as similarity of grayscale or texture. The definition of growth parameters depends on the problem explored and on the image type. In this work the mammograms were segmented based on the gray levels.

In Figure 1(a) the mammogram can be observed. In Figure 1(b) the mammogram after the negative transformation is presented. Figure 1(c) brings the result of the image segmentation of Figure 1(b). In Figure 1(d), the ROI is selected ignoring the rest of the image.

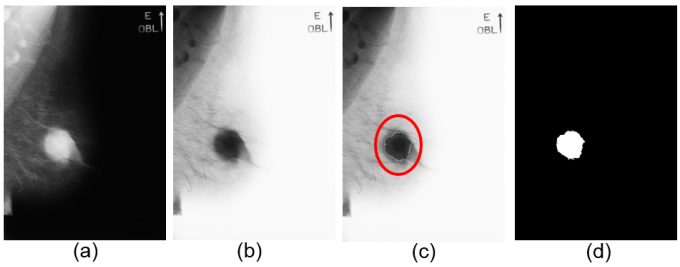

Figure 1. (a) Original image (b) Negative transformation (c) Segmented image (d) ROI

4.3. Feature Extraction

The process of feature extraction allows to obtain relevant characteristics to represent lesions found in mammograms. There are many attributes to represent ROI's on medical images, which are distribuited in three main categories: texture, gradient and geometry (morphology) [Nixon and Aguado 2008].

Texture describes image characteristics such as roughness, uniformity and regularity. The recognition of texture features is a challenge because they do not present a regular pattern and are dependent on scale [Ferreira and Borges 2005]. Texture analysis is usually a very time consuming process, so it is not indicated in scenarios where a large volume of images is involved [Costa et al. 2012]. The gradient is ideal to find changes in gray levels and the direction of edge on image regions [Gonzalez and Woods 2008].

The geometric attributes describe the morphological properties of regions of interest as area, perimeter and circularity. In this work, the morphological attributes were extracted from mammograms due to their high relevance in medical diagnosis to identify objects (regions) of interest [Al-Shamlan and El-Zaart 2010], and because breast tumors are visually distinguishable by their geometric shape.

4.4. Classification

Classification in Image Processing aims to identify characteristics or patterns found in an image and assign them to specific classes [Solomon and Breckon 2011]. When there is previous knowledge about the attributes and their respective classes, the classification is supervised. Otherwise the classification is unsupervised and attributes are grouped according to similarity criteria to form clusters.

Methods of supervised learning such as Decision Tree, Regression Tree, Nearest Neighbour and Artificial Neural Networks are used in this work due to their wide use, especially in the classification of breast tumors. The corresponding algorithms of these methods - J48, CART (Classification And Regression Trees), MLP (Multilayer Perceptron) and IBK - were used in papers cited in Section 2 and also in this work with the objective of comparing the classification performance.

5. Description of the Methodology

Basead on the proposal of [Souto et al. 2014], the methodology implemented in the CAD system, is divided into two phases: training and classification (Figure 2).

The following tasks are performed during the training phase:

- Selection: the mammogram that will be submitted to the training process is selected from an image database. In this phase, the selected image already have a diagnosis;
- Transformation: the image is submitted to negative transformation to highlight the lesions, helping the analyst to better observe the ROIs;
- Segmentation: in this work, the segmentation performed by the region growing algorithm can be considered as semi-automatic and local. The user indicates the initial seed point of the segmentation and empirically sets the Euclidean Distance (ED) according to the characteristics of the lesion.

- Feature extraction: geometric features such as area, perimeter, perimeter-area, shape and fractal (Table 1) from ROI are calculated;
- Creation of the model: features extracted from a set of images and their corresponding diagnoses are used as input for the classification algorithm (J48, CART, MLP or IBK);
- Model: the model characterizes the king of lesion (benign or malignant) in mammograms and is used in the classification phase.

The classification phase is similar to training phase:

- Selection: the mammogram that will be used in the classification process is selected. It does not have a diagnosis in this phase;
- Transformation: the image is submitted to negative transformation to highlight the lesions;
- Segmentation: the identification and segmentation of the ROI are the objectives of this step, similarly to this procedure on the training phase;
- Feature extraction: features (Table 1) are obtained from the ROI;
- Classification: the extracted features are submitted to the created model in the training phase. This classification process indicates the diagnosis of the lesion;
- ROI classified: the result of this phase is the region of interest classified as benign or malignant. This information can help the specialist providing a second opinion about the case or increasing the security of diagnosis.

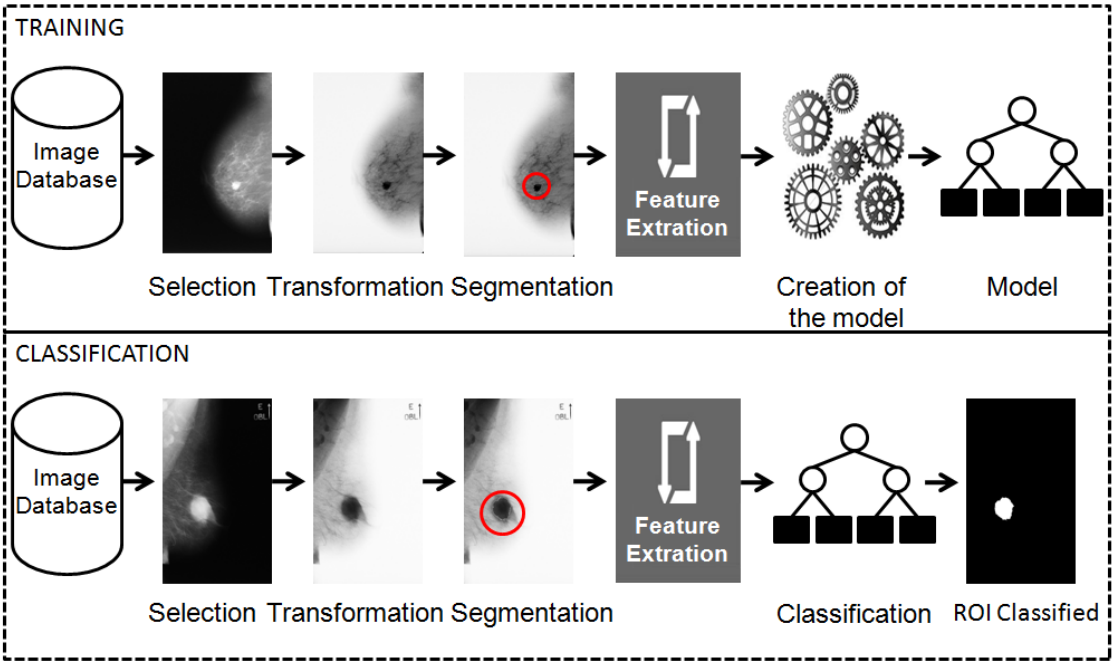

Figure 2. Overview of the methodology

6. Case Study

In this case study, four experiments were carried out using images from the Breast Cancer Digital Repository (BCDR) [BCDR 2018]. The BCDR is a public repository composed of patient cases with ages between 20 and 90 years. For each image there is a manual

segmentation of the lesions and BI-RADS (Breast Imaging Reporting and Data System) classification by specialized radiologists.

The BI-RADS was designed by the American College of Radiology (ACR) to issue unambiguous breast imaging reports assuring a standard language among radiologists, gynecologists and breast clinics [ACR 1993]. The BI-RADS is divided into categories according to the type of lesion found in the image and its probability of malignancy. In this study, only mediolateral oblique (MLO) mammographic views were used and there were no further criteria for selecting images for training and classification phase.

In Experiment 1 the objective was to define the set of most representative morphological attributes of ROIs. The Experiment 2 was performed to evaluate the efficacy of region growing in the segmentation of the ROIs. For this experiment, images in their original format were used, i.e., without the application of preprocessing techniques. Forty mammograms (20 malignant and 20 benign) were segmented by region growing algorithm and the attributes of the ROIs were extracted and submitted to training. For classification, twenty mammograms (10 malignant and 10 benign) were also segmented and had their attributes extracted.

In Experiment 3 the efficiency of the application of negative transformation technique to improve the visualization of lesions in images was evaluated. In this experiment, the same set of images of Experiment 2 was used. In Experiment 4, 70 images for the training phase and 30 for the classification phase were used. The negative transformation technique was applied to images, the segmentation was performed and the set of attributes were extracted to be used as input for data mining algorithms (J48, CART, MLP, IBK), in order to create the training model and then perform the classification of ROI's in 2 classes (benign or malignant).

The results of classification were analyzed using the following statistical measures (the benign term was considered as "negative" and malignant as "positive"): true positive (TP); false positive (FP); true negative (TN); false negative (FN). The TP term refers to cases where the tumor is malignant and the mammogram is correctly classified. In the case of FP the tumor is benign, but is misclassified as malignant. Therefore, the Sensitivity is the capacity to detect true positives, Specificity is the detection of true negatives and Accuracy corresponds to the rate of positive and negative examples correctly classified. The results were calculated using the following equations:

- Sensibility (SS) = $\frac{TP}{(TP+FN)}$
- Specificity (SP) = $\frac{TN}{(TN+FP)}$
- Accuracy (ACC) = $\frac{(TP+TN)}{(TP+TN+FP+FN)}$

7. Results and Discussion

In this section the obtained results in this work are presented and analyzed.

7.1. Experiment 1

Through a bibliographic research, a set of morphological attributes were selected. These features were extract from 40 mammograms and submitted to classification by an algorithm based on Decision Tree, the J48. This technique was applied to select the most

discriminative attributes that increased the information gain in the classification of the lesion as benign or malignant. As a result of this experiment, a set of 12 most representative attributes of regions of interest in mammograms was defined. These attributes were used in all experiments of this work (Table 1).

Table 1. Morphological attributes

Attribute	Equation	Description
Area (A)		Returns the area of the region. Measured in pixels.
Perimeter (P)		Returns the perimeter of the region. Equal to the number of these pixels in the edge of the region.
Fractal	$2\frac{log(0.25*P)}{log(A)}$	Index that measures the shape complexity of the region.
Max radius		Returns the maximum distance between the center and the edge of the region.
Min radius		Returns the minimum distance between the center and the edge of the region.
Circle	$1 - \frac{A}{\pi(radius^2)}$	Returns 0 for circular regions and near of 1 for linear regions.
Circularity	$\sqrt{\frac{MinRadius}{MaxRadius}}$	Measures the similarity of the region with an ellipse.
Compactness	$\left(\frac{2\sqrt{A\pi}}{P}\right)$	Returns the degree of dissimilarity between the region and a perfect circle.
Dispersion	$\frac{MaxRadius}{Area}$	Measures the irregularity of a region.
Shape	$\frac{P}{4\sqrt{A}}$	Returns 1 for compact regions and increases according to the irregularity.
Perimeter-Area	$\frac{P}{A}$	Ratio between the perimeter and the area of a region. It is an indicator of the complexity of the shape of the region.
Spiculation	$Si = \frac{l_i}{b_i^2}$	Ratio between the length of the edge of the region and the square of the width of the region, where l is the length of the edge and b is the base length of the region.

7.2. Experiment 2

In this experiment, the performance of the local and semiautomatic region growing algorithm was analyzed. The RG had inferior performance in segmentation of mammograms with overlapping tumors in dense glandular tissue. This happened because this kind of breast tissue tends to obscure the visibility, masking the lesions. Malignant tumors with irregular shape and indistinct edge also contributed to decrease the performance of the algorithm, because the grayscale pixels of the ROI are very similar to the grayscale pixels of the neighboring regions. On such cases, the variation parameter of the grayscale (Euclidean Distance) was decreased. On cases of benign tumors, which had well-defined shape and usually differ from the rest of the image, there was no difficulty in segmentation process.

In Table 2 are the results of classification with J48, CART, MLP (learning rate = 0.3 and 4 layers of neurons) and IBK (K = 3). It is observed that the classification results were the same for the decision tree algorithms (J48 and CART), which had the highest rates of sensitivity (90%) and accuracy (85%), however, the specificity of both was equal

to 80%. The MLP and IBK algorithms obtained sensibility of 70%, specificity equal to 90% and 80% of accuracy.

Table 2. Classification result after segmentation with region growing

Algorithm	SS (%)	SP (%)	ACC (%)
J48	90	80	85
CART	90	80	85
MLP	70	90	80
IBK	70	90	80

7.3. Experiment 3

In this experiment, the image preprocessing with the negative transformation was performed. This technique does not require that the user inserts parameters, because it consists in reversing the intensity levels of pixels.

The application of negative transformation improved the visualization of lesions, especially in malignant ones, that were much better highlighted in the images. With a better visualization of the lesion, it was possible to choose more clearly the initial seed point and better adjust the ED parameter, resulting in a better segmentation. It reflected in the increase of the classification rates as it is observed in Table 3, which brings the performance of J48, MLP (learning rate = 0.4 and 7 layers of neurons), CART, and IBK (K = 1). This experiment obtained better classification rates than Experiment 2. The algorithms showed sensitivity equal to or greater than 80% and most got up to 90% of specificity.

Table 3. Classification result after negative transformation

Algorithm	SS (%)	SP (%)	ACC (%)
J48	90	90	90
CART	80	90	85
MLP	90	90	90
IBK	90	70	80

7.4. Experiment 4

This experiment had as main objective to evaluate the performance of the classifiers. It was observed that the negative transformation enabled a better visualization of the ROIs, but images of low quality made difficult the visualization and segmentation of ROIs. In these cases, even a doctor or specialist probably wouldn't be able to distinguish the lesion. Therefore, it may be necessary to study and apply other preprocessing techniques to improve the image quality.

In other cases, the RG demonstrated precision in the segmentation of the ROI, resulting in the extraction of representative attributes that contributed to the high rate of classification. Furthermore, the success of the classification is due to the training phase and to the number of samples used, since the classifiers are based on the model to correctly identify the instances in their respective classes.

The final results are shown in Table 4. Using 70 images in the training phase and 30 images for the classification phase, J48, CART and MLP (learning rate = 0.3 and 2 layers of neurons) achieved 100% of sensitivity. The IBK (K = 7) algorithm had the highest specificity rate (100%) and all the classifiers obtained above 90% of accuracy. These results prove the effectiveness of the method.

In Section 2 was shown that Radovic *et al.* (2013) used 20 texture attributes and obtained 79.33% of accuracy in classification [Radovic et al. 2013]. Asad *et al.* (2011), with only 7 attributes extracted from geometric shape of the ROIs, reached the rate of 80% of accuracy [Asad et al. 2011]. Surendiran and Vadivel (2012) used a large number of images for testing and 17 attributes of shape and edge, to reach 93.72% of accuracy [Surendiran and Vadivel 2012]. More recently, Liu and Tang (2014) achieved 94% of accuracy, extracting 12 attributes of shape and texture from 826 images [Liu and Tang 2014].

Comparing those results with this work, which used a set of 12 morphological attributes from 100 images, 96% of accuracy was obtained. This result is due to the application of the negative transformation technique that enhanced the lesions, which facilitated the choice of the starting point for segmentation and the adjustment of the Euclidean Distance. The correct segmentation of the lesion, and the set of morphological attributes defined in this study, contributed to the representation of the ROIs and to the high classification rate.

Table 4. Results of classification

Algorithm	SS (%)	SP (%)	ACC (%)
J48	100	93	96
CART	100	93	96
MLP	100	80	90
IBK	80	100	90

8. Conclusion

In this paper, a methodology and a computer-aided diagnosis system were proposed to assist radiologists, providing a second opinion about the analysis of lesions on mammograms. The new experiments revealed that the negative transformation was useful to enhance visualization of the lesions on mammograms and to facilitate the adjustment of the segmentation parameters. In addition, a set with only 12 morphological attributes was defined to distinguish the benign and maligant lesions.

The classification was performed by J48, CART, IBK and MLP algorithms. The decision tree algorithms, J48 and CART, obtained the best results, allowing the correct classification of 96% of the mammograms, with 100% of sensitivity and 93% of specificity, demonstrating the representativeness of the attributes used and the CAD efficacy. As future work, we intend to increase the number of images for tests, and also to perform the mammogram classification according to BI-RADS, due to its wide use by radiologists and specialists worldwide.

Acknowledgements
The authors acknowledge the financial support of CAPES, CNPq and FAPERN, and also the important contribution of the Mossoro Oncology and Hematology Center for this research.

References

ACR (1993). American college of radiology breast imaging reporting and data system (bi-rads). *Reston, VA: American College of Radiology*.

Al-Shamlan, H. and El-Zaart, A. (2010). Feature extraction values for breast cancer mammography images. *International Conference on Bioinformatics and Biomedical Technology*, pages 335–340.

Alvarenga, A. V., Infantosi, A. F. C., de Azevedo, C. M., and de Albuquerque Pereira, W. C. (2003). Application of morphological operators in the segmentation and determination of the contour of breast tumors in ultrasound images. *Brazilian Journal of Biomedical Engineering*, 19(2):91–101.

Asad, M., Azeemi, N. Z., Zafar, M. F., and Naqvi, S. (2011). Early stage breast cancer detection through mammographic feature analysis. *Bioinformatics and Biomedical Engineering,(iCBBE) 2011 5th International Conference on. IEEE*, pages 1–4.

BCDR (2018). Breast cancer digital repository. https://bcdr.inegi.up.pt.

Calas, M. J. G., Gutfilen, B., and de A. Pereira, W. C. (2012). Cad and mammography: why use this tool? *Brazilian Board of Radiology and Image Diagnosis*, 1:46–52.

Costa, A. F., Humpire-Mamani, G., and Traina, A. J. M. (2012). An efficient algorithm for fractal analysis of textures. *SIBGRAPI*, 25th Brazilian Symposium on Computer Graphics and Image Processing:39–46.

Cuadros, O., Botelho, G., Rodrigues, F., and Neto, J. B. (2012). Segmentation of large images with complex networks. *SIBGRAPI*, 25th Brazilian Symposium on Computer Graphics and Image Processing:24–31.

dos Santos Romualdo, L. C., da Costa Vieira, M. A., and Schiabel, H. (2009). Mammography images restoration by quantum noise reduction and inverse mtf filtering. *SIBGRAPI*, 22th Brazilian Symposium on Computer Graphics and Image Processing:180–185.

Ferreira, C. B. R. and Borges, D. L. (2005). A selection strategy of a minimal subset of wavelet features in a multiresolution approach for the classification of tumors on mammograms. *V Workshop of Medical Informatics*, 1.

Gonzalez, R. C. and Woods, R. E. (2008). *Digital Image Processing*. Prentice Hall, 5 edition.

Leite, G. C., Leite, J. S. S., Meneses, F. G. A., Santos, D. A., and Silva, J. S. (2011). The use of thresholding techniques to aid in the diagnosis of breast cancer. *III National Encounter Of Biomechanical Engineering*.

Liu, X. and Tang, J. (2014). Mass classification in mammograms using selected geometry and texture features, and a new svm-based feature selection method. *IEEE Systems Journal*, 8(3):910–920.

Nixon, M. S. and Aguado, A. S. (2008). *Feature Extraction & Image Processing*. Academic Press, 2 edition.

Radovic, M., Djokovic, M., Peulic, A., and Filipovic, N. (2013). Application of data mining algorithms for mammogram classification. *Bioinformatics and Bioengineering (BIBE), 2013 IEEE 13th International Conference on. IEEE*, pages 1–4.

Skaane, P., Engedal, K., and Skjennald, A. (1997). Interobserver variation in the interpretation of breast imaging. *Acta Radiol*, pages 497–502.

Solomon, C. and Breckon, T. (2011). *Fundamentals of digital image processing: a practical approach with examples in matlab*. Wiley-Blackwell.

Souto, L. P. M., dos Santos, T. K. L., and dos Santos Silva, M. P. (2014). Métricas morfológicas para a classificação de tumores de mama. *Workshop de Informática Médica. CSBC*.

Surendiran, B. and Vadivel, A. (2012). Mammogram mass classification using various geometric shape and margin features for early detection of breast cancer. *Int. J. Medical Engineering and Informatics*, 4(1):36–54.

Thurfjell, E., Lernevall, K., and Taube, A. (1994). Benefit of independent double reading in a populationbased mammography screening program. *Radiology*, pages 241–244.