Novel approach to classification of Abnormalities in the mammogram image

Khalid El Fahssi,Abdelali Elmoufidi,Abdenbi Abenaou, Said Jai-Andaloussi, Abderrahim Sekkaki

***Abstract*—Mammography remains the most effective tool for the early detection of breast cancer and Computer Aided Diagnosis (CAD) is usually used as a second opinion by the radiologists. segmentation and classification of breast masses in mammography play a crucial role in Computer Aided Diagnosis system (CAD) . In this paper we propose an approach based on the theory of adaptive orthogonal transformation that will calculate the informative characteristics of regions of interest of mammography images and classification by comparison of the similarity between the vectors of the characteristics of regions of interest by use of the coefficient of matrix of correlation. The result obtained by this calculation method allows the increase the efficiency of diagnosis. To illustrate the effectiveness of the method we present the results of experiments carried out on the basis of images MIAS mammograms.**

***Keywords*—segmentation; mammography; active contours; classification;orthogonal;correlation**

I. Introduction

Breast cancer is the most common evil in women worldwide. It is a leading cause of female mortality [1, 2,3] since every woman has a one in eight chance of developing breast cancer during her lifetime. [4]. The Prevention of the disease is very difficult because the risk factors are either poorly known or not suggestible. Scientific studies have provided a better understanding of development of cancer but it is still not possible to know why one person develops breast cancer. It should be noted that only 5 to 10% of breast cancers are hereditary related to the transmission of deleterious genes which are the most frequently incriminated are BRCA1 and BRCA2 (breast Cancer acronyms) associated with a predisposition of the disease [5] . Mammography is the most effective imaging technique to detect tumors at an early stage [2,6] and currently is the principal investigation in breast cancer screening. However, radiologists are always obliged to examine very finely the image to a better diagnosis hence the importance of Computer Aided Diagnosis system (CAD) developed over the last twenty years[7]. Generally systems (CAD) based on a series of approaches including pretreatment steps, segmentation, extraction of parameters of classification and finally the interpretation and classification of suspected abnormalities [7,8,9,10] . Generally the first indicators of malignancy parameters are connected to the mass density, size, shape and edges, malignant masses often infiltrate the fabric of linear son purposes that extend outward from the center of the mass. From the pathology of malignant masses, the shape can be used to discriminate between malignant mass and benign masses. Approximately 80 to 85% of breast cancers are diagnosed localized tumor's appearance on mammography[4]

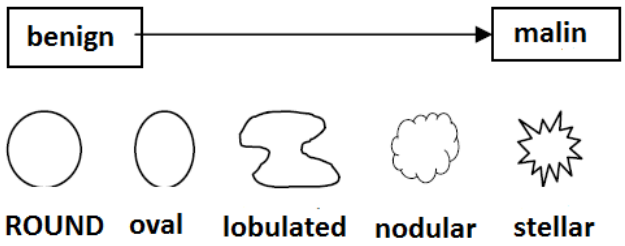


Fig. 1. The morphological spectrum of mammographic masses.

For automatic classification of tumors several techniques are used such as that based on artificial neural networks [11,12], one based on fuzzy logic [14], and other based on support vector machines [13,26]. In this paper we propose a new technique based on adaptive orthogonal transformation and the coefficient of correlation matrix.
Figure 2 define the different step of our system of detection and classification of lesions.

Khalid El Fahssi,Abdelali Elmoufidi,Abdenbi Abenaou, Said Jai-Andaloussi, Abderrahim Sekkaki are with LIAD Labs, Casablanca, Kingdom of Morocco.; elfahssi@etude.univcasa.ma

Khalid El Fahssi,El Moufdi, Said Jai-Andaloussi and Abderrahim Sekkaki are with Faculty of science Ain-chok, Casablanca, Kingdom of Morocco.

Abdenbi Abenaou are with National School of Applied Sciences, Agadir, Kingdom of Morocco.

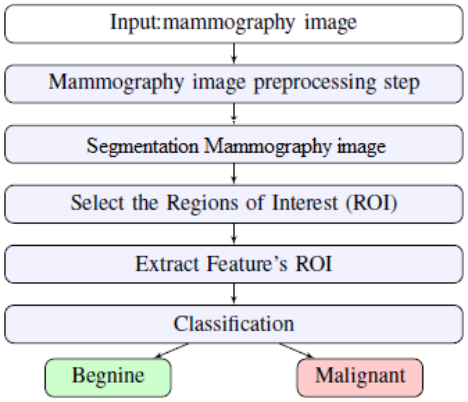


Fig. 2. Organization chart of the proposed method

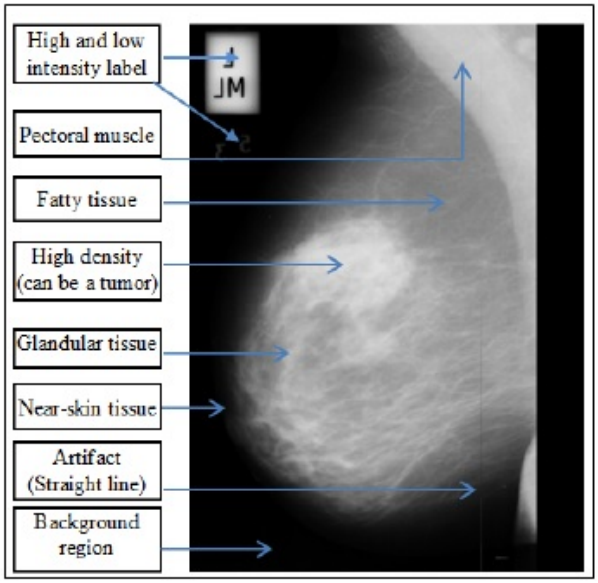


Fig. 4. The different components in the mammogram image

This paper is organized as follows: Section II.A describes the database used for evaluation, section II.B we present the different classification methods ,section II.C we give our method of pretreatment, section II.D we give a result of our method of segmentation, section II.E present the method for classification and detection of the lesion. The result is presented in section III and we end with a conclusion and perspective in section IV.

II. Materials and Method

A. Database

The mini-MIAS [20] database contains a total of data 322 MLO view mammography images. This database is divided into categories such margin: speculated, circumscribed or poorly defined. The images have a resolution of 1024x1024 pixels. From this data set, a total of 111 lesions was selected. These include 60 benign and 51 malignant masses. An example of a series of the image is given by Figure 3 and different components in the mammogram by Figure 4.

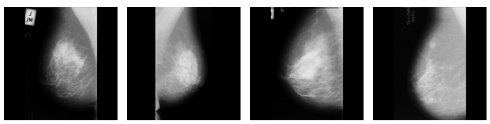


Fig. 3. Example of mammography study.

B. Classification Methods

the notion of Classification means assigning a label to samples of a database using a number of characteristics. These features must obviously be able to identify each sample. In image processing, the echantillon may designate a pixel, an area in the image, an object represented in the image or the image itself. Depending on the application, the purpose of the classification is either:

- Classify the pixels of the image into different zones. In this case, the classification problem reduces to a problem of image segmentation into different objects. For example, one can classify the different areas of a mammographic in lesion or non-lesion.

- Classify the image or objects an the image in different categories. For example, we can classify the masses who are in mammographic images in malignant or benign.

Two methods of classification can be distinguished: the unsupervised classifications and those supervised.

1) Unsupervised classification methods: These techniques are used when the identity of classes is not known. This results from a lack of information from the study population. There

are classification algorithms composed of several iterations, to create groups of individuals with similar characters. The unsupervised classification, called automatic or clustering consists to identify the different classes naturally without any prior knowledge. The objective, in this case, is to identify a structure in the images of the database based on their content. The images are assigned to different classes estimated using two essential criteria are the great homogeneity of each class and good separation between classes. Among the unsupervised classification methods the most commonly used is that of the K-means algorithm, also called dynamic clustering algorithm (McQueen, 1967). The algorithm works by specifying the number of K classes (clusters) expected (K being set by the user). It calculates the intra-class distance and reattached cluster centers according to distance values. The disadvantages of this method are firstly the need to set the number of classes before starting the classification. Secondly, this method is very sensitive to the initial data distribution. Finally, this method assumes that classes follow the reduced normal distribution laws, in other words, with the same importance in all directions of space which is not always verified.

Another method of clustering is the self-organizing map (SOM) (Kohonen, 1984). SOM is a neural network, supervised by a non-competitive process, is capable of projecting large data in a two-dimensional space. During learning, each neuron is specialized in the recognition of a certain type of input. The self-organizing map consists of a set of neurons interconnected. A configuration between the inlet space and the space of the network is built, and, in two points near the input space near activate two units on the map. This method is more robust to initial conditions that the K-means algorithm. The major drawback of this method is the CPU time associated with iterations allowing the construction of the self-organizing map.

2) Supervised classification methods: If the user has sufficient information on the study population (as is the case of breast images), it can perform a supervised classification. This category is supposed to have a group of individuals of each class, which we know they belong. These individuals form d'apprentissage samples. They are used to train the classifier. Other samples, called de test serve to validate the classification by assessing its relevance through the rate of correctly classified individuals. There are several supervised classification methods. The most famous methods are the linear discriminant analysis, logistic regression, neural networks. Some research has focused on the linear discriminant analysis. This is a simple method of classification between images belonging to different classes based on linear analysis. The main idea of this technique is to build decision limits directly optimizing the error criterion. However, this method is suited to data linearly separable which is not always the case. Artificial neural networks (ANN), are widely used for classification problems. They are based on the theory of perceptrons. ANN consists of several neurons distributed over an input layer (denoting the descriptors), an output layer (designating the classification result) and a number of hidden layers. Moreover, this method is capable of modeling non-linear systems very complex. However, the disadvantage of this method is the choice of the number of hidden layers and the number of neurons in each layer. Thus, the user is brought to experiment with different combinations of the number of layers and neurons in order to arrive at the most appropriate' neural network has its type of application. By cons, the neural networks of radial basis functions (RBF) are constituted by a single hidden layer. The major advantage over other artificial neural networks is the use of a less complex structure (only one hidden layer). In addition, the computational complexity induced learning is less than that induced by learning ANN.

through to the existence of hybrid algorithms. However, performance of such a network dependent for a selection of basic functions, the number of functions constituting the radial basis function (number of unit'es the hidden layer) and estimation of network parameters . Other research has been directed towards the logistic regression (LR). This is a multivariate model commonly used in epidemiology (or cancer). It is used when the output variable (the level) is qualitative, usually binary (the occurrence or absence of a disease). The input variables (descriptors) can be against by either qualitative or quantitative. Logistic regression is able to achieve a probability estimate using logistic formulation. Faced with linear functions, the separators wide margin, known by the acronym SVM (Support Vector Machines) are originally designed for binary classification problems. They allow to linearly separate the positive examples from the negative examples in the set of training images by a hyper-plane that guarantees maximum margin (Vapnik, 1999). The effectiveness of SVMs is often greater than that of all other supervised classification methods. For problems of non-eparabilit, SVMs used to perform a nonlinear transformation of the input observations in a higher dimensional space

to reduce to the linear case. In addition, SVMs can also address the multi-class classification problems.

C. preprocessing of mammography Image

This step is crucial in our system for the reduction of false positives. Indeed, in these images, the breast covers only 30% of the entire image [15]. Moreover, the digitization of mammography images may result in noise in the resulting mammograms. So in mammography images several types of noise and artifacts are present [16]. Thus for improving the quality of mammography we used an image preprocessing method based on shrinkwrap function [17], this function amplifies areas of high intensity and segments them using a front. The front is initialized on the convex hull (for speed) and erodes the map until it has converged on the edge of the areas to keep, maintaining edge geometry.

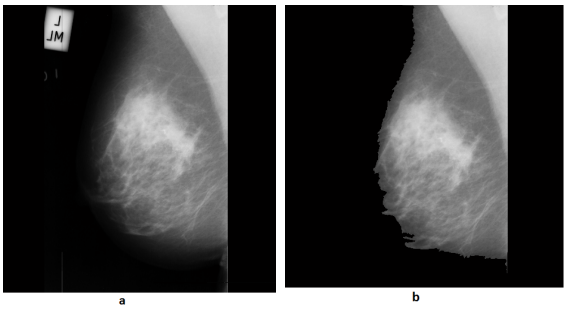


Fig. 5. Mammogram preprocessing step : (a) Original mammogram, (b) Artifact suppressed

D. our method of segmentation

For segmentation the mammography images we have used a combined solution of the two approaches, one based on levels set theory and the other based on the principle of the minimization of the energy of active contours[18]. The elaborated algorithm which is an interactive approach permits to segment effectively the images from its coarse resolution to its refinement. The results of the segmentation are illustrated from the database of MIAS. The figure represents the result of our method of segmentation.

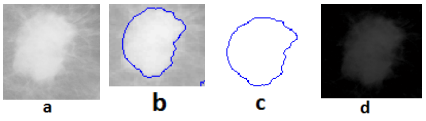


Fig. 6. Segmentation of mammograms mdb184 : (a) Original mammogram, (b)ROI detected ,(c) The contour of ROI, (d) ROI extracted

E. Method proposed for classification of lesion

The first part of the approach proposed in this article has allowed the segmentation and selection of regions of interest of mammography images of various types of disease. The classification of mammograms depending on the type of disease images requires the extraction of the most relevant features in these regions of interest [19],[21].

the calculation of characteristics vectors informative regions is ensured by the use of an adaptable orthogonal transformation [22]for each class regions (type of disease). The spectrum of informative characteristics of a given area (class) obtained decreases rapidly. In other words, the analysis of the spectra obtained allows a considerable difference between the classes of regions. This property will subsequently develop a simple decision rule which ensured a high certainty of classification.

The principle of the proposed method is to synthesize an adaptable operator orthogonal transformation to generate functions of bases parameterizable. Using these transformations [22],[23] is favored by the ability to adapt to the shape of their basic functions depending on the nature of the standard vector formed by mammograms each class images. In other words, for each type of disease a system of basis functions are associated parameterizable for showing his images. In addition, these functions of base parameterizable meet the criteria for completeness of the system ensuring the transformation of vectors without loss of information content. The system of basis functions formed is expressed as a factorisable orthonormal matrix operator, which therefore allows a transformation with a fast calculation algorithm:

$$Y = \frac{1}{N} HX \quad (15)$$

where:
- $X = [x_1, x_2, ..., x_N]^T$ is the vector representing the region of interest in the segmented image mammography given initial vector transform (of size $N = 2^n$).
- $Y = [y_1, y_2, ..., y_N]^T$ is the vector of informational characteristics calculated by the spectral operator orthogonal H of dimension N x N.
Factorization of Good [24] showed a possibility of representing the matrix operator H as product G_i (16) Sparse matrix with a higher proportion of zero which has allowed the construction the quick transformation algorithms of Fourier , Haar and Walsh. The matrices $G_i (i = 1, ..., n)$ are constructed by blocks of matrices $V_{i,j}$ of minimum dimension that is called spectral nuclei [22]:

$$G_i = \begin{bmatrix} \begin{bmatrix} \alpha_{i1} & 0 \\ \beta_{i1} & 0 \end{bmatrix} & \cdots & 0 & \begin{bmatrix} \gamma_{i1} & 0 \\ \delta_{i1} & 0 \end{bmatrix} & \cdots & 0 \\ 0 & \begin{bmatrix} \alpha_{i2} & 0 \\ \beta_{i2} & 0 \end{bmatrix} & & 0 & \begin{bmatrix} \gamma_{i2} & 0 \\ \delta_{i2} & 0 \end{bmatrix} & & 0 \\ \vdots & \ddots & & \vdots & \ddots & & \vdots \\ 0 & \cdots & 0 & \begin{bmatrix} \alpha_{i,N/2} & 0 \\ \beta_{i,N/2} & 0 \end{bmatrix} & \cdots & 0 & \begin{bmatrix} \gamma_{i,N/2} & 0 \\ \delta_{i,N/2} & 0 \end{bmatrix} \end{bmatrix} \quad (16)$$

With

$$V_{i,j} = \begin{bmatrix} \alpha_{ij} & \cdots & \gamma_{ij} \\ \beta_{ij} & \cdots & \delta_{ij} \end{bmatrix} = \begin{bmatrix} \cos(\varphi_{i,j}) & \cdots & w_{i,j}\sin(\varphi_{i,j}) \\ \sin(\varphi_{i,j}) & \cdots & -w_{i,j}\cos(\varphi_{i,j}) \end{bmatrix},$$
$$w_{i,j} = \exp(j\theta_{i,j}), \quad \varphi \in [0, 2\pi], \quad \theta \in [0, 2\pi].$$

Hence equation (15) can be written as follows:

$$Y = \frac{1}{N} HX = \frac{1}{N} G_1 G_2 .... G_N X = \frac{1}{N} \Pi_{i=1}^{N} G_i \quad (17)$$

By defining the parameters $\varphi_{i,j}$ and $\theta_{i,j}$ can train operators orthogonal of transformations with basic functions complex , and $\theta_{i,j} = 0$ operators with real functions. Adapting Operator H in (15) is provided by the condition:

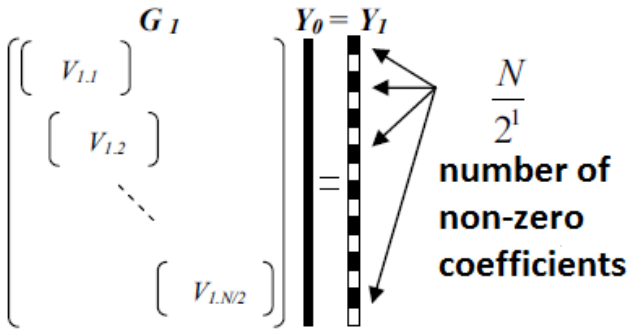
$$\frac{1}{N} H_a Z_{sd} = Y_c = [Y_{c,1}, 0, 0, ..., 0], Y_{c,1} \neq 0 \quad (18)$$

where:
- Y_c is the target that builds the basis for adjusting the vector operator H_a.
- Z_{sd} represents the calculated by means of the estimates of the statistical characteristics of images of a given class standard vector.
- H_a is adaptable to synthesize operator.

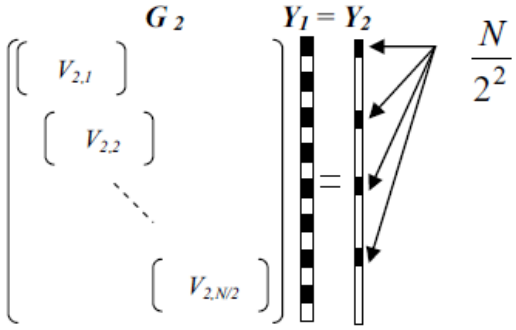
Synthesis adaptable operator H_a based standard Z_{sd} (for a given class) is to calculate the angular parameters $\varphi_{i,j}$ matrices G_i according to condition (18). The procedure for calculation of the parameters is based on an iterative algorithm to calculate the target vector Y_c by step according to the equation:

$$Y_i = G_i Y_{(i-1)} \quad (19)$$

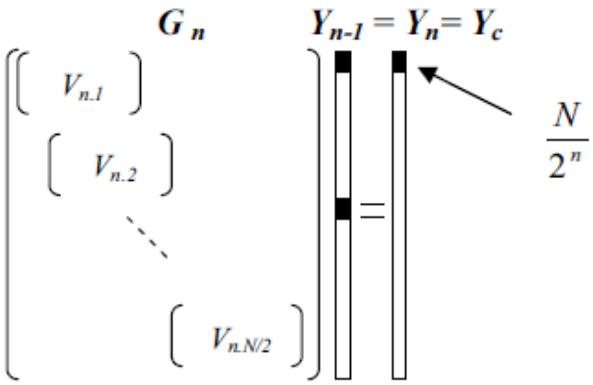
The first iteration $(Y0 = Z_{sd})$



The second iteration



The nth iteration



The final calculation of the vector Y_c allows obtaining adaptable operator H_a.

For the classification of lesions, we dispose two sets of feature vectors for each type of tumor (benign or malignant). The first was used for the calculation of the standard $Z_{sd,i}$ (tumor i) for the synthesis of the operator, while the second set used to form the spectral standard $Y_{sd,i}$. The latter is obtained by projecting the feature vector of the second set in Ha adaptable bases.
To make the decision and classify each tumor, we calculate each Y_i spectrum in each base $H_{a,i}$ of the region of interest, then we sills of a regle of decision based on the calculation of the coefficient of matrix correlation between the standard spectrum of each class and the projection of Y_i spectrum of the region of interest in the base $H_{a,i}$ of the same class. Correlation coefficient is calculated by the following equation:

$$r_p = \frac{\hat{\sigma}_{XY}}{\hat{\sigma}_X \hat{\sigma}_Y}$$

with

$$\hat{\sigma}_{XY} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x}) \cdot (y_i - \bar{y})$$

$$\hat{\sigma}_X = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2}, \hat{\sigma}_Y = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \ , \ \bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$$

which are respectively estimators of covariance, standard deviations and expectations of the variables X and Y.
Then two spectra are not linearly correlated if r_p is zero. The two spectrum are better correlated if r_p is near to 1 or -1. then according to the condition of linear correlation, the specter of the region of interest belongs to a class of which $max(|r_p|)$ is near to 1.

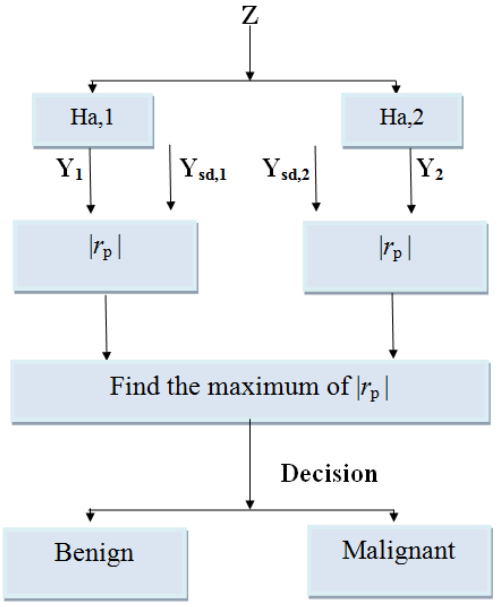


Fig. 7. Classification procedure

III. Results and Discussion

This section presents the results of experiments performed on 111 selected mammography database mini-MIAS images. To evaluate the effectiveness of the proposed method, we used the information provided by the Mammographic Image Analysis Society database (MIAS) including the class of anomalous images and coordinates of their centers of regions of interest. To test our method, the algorithm is applied to each image containing a mass lesion. In the classification of ROI to Benign or Malignant, a positive case means correct classification of ROI to benign or malignant while a negative case means incorrect classification of ROI as such a type. The definitions of the fractions are as below:

-True Positive (TP) means breast classified as Malignant that proved to be Malignant.

- False Positive (FP) means breast classified as Malignant that proved to be Benign.

- False Negative (FN) means breast classified as Benign that proved to be Malignant.

- True Negative (TN) means breast classified as Benign that proved to be Benign.

We have tested the performance of our method by calculating and analysis of accuracy, sensitivity and specificity for malignant and benign

classification. These are defined and calculated as follows[25]:

Sensitivity: number of correct classified Benign mass/number of total Benign mass :

$$Sensitivity = (\frac{TP}{TP + FN}) * 100\%$$

Specificity: number of correct classified Malignant mass/number of total Malignant mass:

$$Specificity = (\frac{TN}{TN + FP}) * 100\%$$

Accuracy: number of correct classified mass/number of total mass:

$$Accuracy = (\frac{TP + TN}{TP + FN + TN + FP}) * 100\%$$

Out method has been proven effective for the classification masses to benign or malignant on a mammogram with a sensitivity equal 93.78%, Specificity equal 94.54% and accuracy equal 93.89%.

IV. Conclusion and Perspective

In this work we presented an approach for classification of lesion to benign or malignant in digital mammograms. The experimentation gives a percentage of 93.78% of a sensitivity and a percentage of 94.54% of Specificity and accuracy equal 93.89% for all cases studied. The results of the algorithm can contribute to solving the main problem in mammography image processing such as diagnostic and classification. The Efficiency of the proposed method confirms the possibility of its use in improving the computer-aided diagnosis.

References:

[1] Al Mutaz M. Abdalla, Safaai Dress, Nazar Zaki , Detection of Masses in Digital Mammogram Using Second Order Statistics and Artificial Neural Network, International Journal of Computer Science & Information Technology (IJCSIT), Vol 3, No 3,June 2011, pp. 176-186.

[2] Liyakathunisa & C.N. Ravi Kumar, A Novel and Efficient Lifting Scheme based Super Resolution Reconstruction for Early Detection of Cancer in Low Resolution Mammogram Images ,International Journal of Biometrics and Bioinformatics (IJBB), Volume (5) : Issue (2) : 2011 , pp 53-75.

[3] N.Ben Hamad, N. Benromdhane, K. Taouil, M.S. Bouhlel, Reduction the False Positives in Assistance Systems Diagnosis in the Breast Cancer, SETIT 2007 4th International Conference: Sciences of Electronic, Technologies of Information and Telecommunications, Tunisia ,2007.

[4] L. M. Bruce & al,1999 "Classifying Mammographic Mass Shapes Using the Wavelet Transform Modulus Maxima Method" IEEE Transactions On Medical Imaging Vol 18 No. 12.

[5] Gulsrud, Thor. O & al, 1996"Optimal filter for detection of stellate lesions and circumscribed masses in mammograms",Poc. SPIE vol 2727, pp 430-440, Visual communications and Image Processing

[6] Songyang Yu and Ling Guan, A Cad System for the automatic detection of clustered microcalcifications in digitized mammogram films, ieee on medical imaging, vol. 19, n. 2, february 2000, pp. 115-126

[7] Rangaraj M.Rangayyan, Fabio J.Ayres, J.E.Leo Desautels, A review of computer-aided diagnosis of breast cancer: toward the detection of subtle signs, sciencesdirect , journal of the franklin institute 344 (2007) 312348

[8] J. B. Jona, N. Nagaveni A Hybrid Swarm Optimization approach for Feature set reduction in Digital Mammograms, WSEAS TRANSACTIONS on INFORMATION SCIENCE and APPLICATIONS, Issue 11, Volume 9, November 2012.

[9] T.Balakumaran, Dr.Ila.Vennila, C.Gowri Shankar, Detection of microcalcification in mammograms using wavelet transform and fuzzy shell clustering, (ijcsis) international journal of computer science and information security,vol. 7, no. 1, 2010

[10] Arianna Mencattini, Marcello Salmeri, Giulia Rabottino, Simona Salicone, Metrological characterization of a cadx system forthe classification of breast masses in mammograms , ieee transactions on instrumentation and measurement, vol. 59, no. 11, november 2010.

[11] Nguyen, H.; Hung, W.T.; Thornton, B.S.; Thornton, E.; Lee, W."Classification of microcalcifications in mammograms using artificial neural networks",Proceedings of the 20th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 1998.

[12] Jasmine, J.S.L.; Govardhan, A.; Baskaran, S."Microcalcification detection in digital mammograms based on wavelet analysis and neural networks",International Conference on Control, Automation, Communication and Energy Conservation, 2009. INCACEC 2009.

[13] Aarthi, R.; Divya, K.; Komala, N.; Kavitha, S."Application of Feature Extraction and clustering in mammogram classification using Support Vector Machine",2011 Third International Conference on Advanced Computing (ICoAC),

[14] Chumklin, S.; Auephanwiriyakul, S.; Theera-Umpon, N."Microcalcification detection in mammograms using interval type-2 fuzzy logic system with automatic member-

ship function generation",2010 IEEE International Conference on Fuzzy Systems (FUZZ).

[15] Songyang Yu and Ling Guan, A CAD System for the Automatic Detection of Clustered Microcalcifications in Digitized Mammogram Films, IEEE Tran. on Medical Imaging, vol. 19, n. 2, February 2000, pp. 115-126.

[16] Stylianos.D et al., A fully automated scheme for mammographic segmentation and classification based on breast density and asymmetry, computer methods and programs in biomedicine 2011, 47-63.

[17] C.W.A.M. van Overveld (Department of Mathematics and Computing Science) and B.Wyvill (Department of Computer Science University of Calgary),"Shrinkwrap: An ecient adaptive algorithm for triangulating an isosurface".

[18] K.El Fahssi, A.Eloufidi et al., Mass segmentation in mammograms based on the minimisation of energy and active contour model, IEEE International Symposium on Medical Measurements and Applications (MeMeA14), june 11-12, 2014, LISBON, PORTUGAL.

[19] Miheala lascu, Dan Lascu Feature Extraction in Digital Mammography Using LabVIEW,2005 WSEAS Int. Conf. on DYNAMICAL SYSTEMS and CONTROL, Venice, Italy, November 2-4, 2005 (pp427-432).

[20] http://peipa.essex.ac.uk/info/mias.html

[21] Osslan Osiris Vergara Villegas,Humberto de Jesus Ochoa Dominguez, Vianey Goadalupe Cruz Sanchez,Efren David Gutierrez Casas, Gerardo Reyes Salgado Rules and Feature Extraction for Microcalcifications Detection in Digital Mammograms Using Neuro-Symbolic Hybrid Systems and Undecimated Filter Banks ,WSEAS TRANSACTIONS on SIGNAL PROCESSING, ssue 8, Volume 4, August 2008.

[22] Abenaou A., Sadik M. Elaboration d une methode de compression des signaux aleatoires base d une transformation orthogonale parametrable avec algorithme rapide WOTIC2011, ENSEM,Casablanca.

[23] Abenaou A. Sadik M. Mthode et algorithme de formation dun systme de fonctions de base adaptables pour le diagnostic des signaux biologiques, Colloque International des Tlcommunications, Tanger 2011.

[24] Good, I.J., The interaction algorithm and practical Fourier analysis, J. Roy. Statist. Soc. Ser. B, B-20, 361-372, 1958, B-22, 372-375, 1960.

[25] R.Mousa and al., Breast cancer diagnosis system based on wavelet analysis and fuzzy-neural, Expert Systems with Applications 28 (2005) 713723.

[26] Abdelali Elmoufid and al,"Automatic Diagnosis and Classification of Abnormalities in Digital X-ray Mammograms",WSEAS TRANSACTIONS on BIOLOGY and BIOMEDICINE,Volume 13, 2016.