Classifying Benign and Malignant Mass using GLCM and GLRLM based Texture Features from Mammogram

Aswini Kumar Mohanty*, Swapnasikta Beberta**, Saroj Kumar Lenka***

*(Phd. Scholar, SOA University,Bhubaneswar, Orissa, India)
**(M.Tech. Scholar,BPUT.Rourkela,Orissa,India)
*** (Department Of Computer Science, Modi Univesity, Lakshmangarh-332311 Rajasthan, India)

ABSTRACT

Mammogram–breast x-ray is considered the most effective, low cost, and reliable method in early detection of breast cancer. Although general rules for the differentiation between benign and malignant breast lesion exist, only 15 to 30% of masses referred for surgical biopsy are actually malignant. In this work, an approach is proposed to develop a computer-aided classification system for cancer detection from digital mammograms. The proposed system consists of three major steps. The first step is region of interest (ROI) extraction of 256×256 pixels size. The second step is the feature extraction; we used a set of 19 GLCM and GLRLM features and the 19 (nineteen) features extracted from grey level run-length matrix and grey-level co-occurrence matrix could distinguishing malignant masses from benign mass with an accuracy 94.9%.Further analysis carried out by involving only 12 of the 19 features extracted, which consists of 5 features extracted from GLCM matrix and 7 features extracted from GLRL matrix. The 12 selected features are: Energy, Inertia, Entropy, Maxprob, Inverse, SRE, LRE, GLN, RLN, LGRE, HGRE, and SRLGE, ARM with 12 features as prediction can distinguish malignant mass image and benign mass with a level of accuracy of 92.3%. Further analysis showing that Area Under the Receiver Operating Curve was 0.995, which means that the accuracy level of classification is good or very good. Based on that data, it concluded that texture analysis based on GLCM and GLRLM could distinguish malignant image and benign image with considerably good result. The third step is the classification process; we used the technique of association rule mining using image content to classify between normal and cancerous mass. The proposed system was shown to have the large potential for cancer detection from digital mammograms

Key words:-Gray-level **Co-Occurrence Matrix, Gray-level Run Length Matrix, mammograms, benign mass, malignant mass, texture features, textures analysis, association rule mining, Receiver operating characteristics.**

I. INTRODUCTION

Breast Cancer is one of the most common cancers, leading to cause of death among women, especially in developed countries. There is no primary prevention since cause is still not understood. So, early detection of the stage of cancer allows treatment which could lead to high survival rate. Mammography is currently the most effective imaging modality for breast cancer screening. However, 10-30% of breast cancers are missed at mammography [1]. Mining information and knowledge from large database has been recognized by many researchers as a key research topic in database system and machine learning Researches that use data mining approach in image learning can be found in [2-8].

Data mining of medical images is used to collect effective models, relations, rules, abnormalities and patterns from large volume of data. This procedure can accelerate the diagnosis process and decision-making. Different methods of data mining have been used to detect and classify anomalies in mammogram images such as wavelets [9,10], statistical methods and most of them used feature extracted using image processing techniques [5].Some other methods are based on fuzzy theory [1] and neural networks [11].

In this paper we have used classification method called Association rule classifier for image classification and the process typically involves two phases: training phase and testing phase. In training phase the properties of typical image features are isolated and based on this training class is created .In the subsequent testing phase , these feature space partitions are used to classify the image. We have used ARM method [12,13] by extracting low level image features for classification. The merits of this method are effective feature extraction, selection and efficient classification. The rest of the paper is organized as follows. Section 2 presents the material and methods, preprocessing, feature extraction and section 3 presents the classification phase. Section 4 discusses the result and section 5 discusses the conclusion.

II. MATERIALS AND METHOD

This study is done through two main phases; the learning phase and the testing phase. Through the learning phase, the system how to differentiate between normal and cancerous cases is learned by using normal and cancerous images. In the testing phase, the performance of the system is test by entering a test image to compute the correctness degree of the system decision.

2.1 MAMMOGRAM DATABASE

The mammogram images used in this paper are provided by the University of South Florida, the digital database for screening mammography (DDSM) [14]. The dataset consists of digitized mammogram images, composed of both oblique and cranio-caudal views. Each mammogram shows one or more tumor mass marked by expert radiologists. The position of individual masses is marked. The location of the abnormalities in form of its boundary provided as chain code where the first two values are the starting column and row of the lesion boundary while other numbers correspond to a specific direction on the X and Y coordinates. The images are digitized from films using the Lumysis scanner with 12 bits depth Maintaining the Integrity of the Specifications

2.2. PRE-PROCESSING

The mammogram image for this study is taken from Mammography Image Analysis Society (MIAS), which is an UK research group organization related to the Breast cancer investigation. As mammograms are difficult to interpret, preprocessing is necessary to improve the quality of image and make the feature extraction phase as an easier and reliable one [15]. The calcification cluster/tumor is surrounded by breast tissue that masks the calcifications preventing accurate detection and shown in Figures.1 .A pre-processing; usually noise-reducing step is applied to improve image and calcification contrast.

In this work an efficient filter referred to as the low pass filter, was applied to the image that maintained calcifications while suppressing unimportant image features. Figures 2 shows representative output image of the filter for a image cluster in figure 1. By comparing the two images, we observe background mammography structures are removed while calcifications are preserved. This simplifies the further tumor detection step.

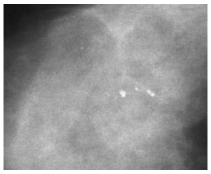


Fig.1 ROI of a Benign

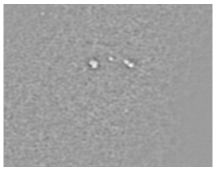


Fig. 2 ROI after Pre-processing

2.3. SELECTION OF ROI

Using the contour supplied by the DDSM for each mammogram [16-19], the ROI of size 50×50 pixels is extracted with mass centered in the window, and divided into two sets: the learning set and the testing set. The learning set is composed of 88 cancerous images and 88 normal images while the testing set contained 23 malignant images and 55 benign images. The benign images are taken from the same image that has malignant regions

2.4. FEATURE EXTRACTION

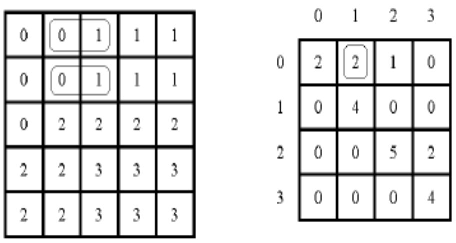
A typical mammogram contains a vast amount of heterogeneous information that depicts different tissues, vessels, ducts, chest skin, breast edge, the film, and the X-ray machine characteristics. In order to build a robust diagnostic system towards correctly classifying normal and abnormal regions of mammograms [20], we have to present all the available information that exists in mammograms to the diagnostic system so that it can easily discriminate between the normal and the abnormal tissue. However, the use of all the heterogeneous information, results to high dimensioned feature vectors that degrade the diagnostic accuracy of the utilized systems significantly as well as increase their computational complexity. Therefore, reliable feature vectors should be considered that reduce the amount of irrelevant information thus producing robust Mammographic descriptors of compact size. In our approach, we examined a set of 19 features were applied to the ROI using a window of size 50 pixels with 50 pixels shift, i.e. no overlap.

The features extracted in this study divided into two categories: grey-level co-occurrence matrix (GLCM) and the other extraction is based on grey-level run-length matrix (GLRLM)[21-25].

2.5. Grey-Level Co-Occurrence Matrix

In a statistical texture analysis, texture features were computed on the basis of statistical distribution of pixel intensity at a given position relative to others in a matrix of pixel representing image. Depending on the number of pixels or dots in each combination, we have the first-order statistics, second-order statistics or higher-order statistics. Feature extraction based on grey-level co-occurrence matrix (GLCM) is the second-order statistics that can be use to analysing image as a texture (Albregtsen, 1995:1). GLCM (also called gray tone spatial dependency matrix) is a tabulation of the frequencies or how often a combination of pixel brightness values in an image occurs (Hall-Beyer, 2005).

The figure below represents the formation of the GLCM of the grey-level (4 levels) image at the distance d = 1 and the direction of 0°.



0	0	1	1	1
0	0	1	1	1
0	2	2	2	2
2	2	3	3	3
2	2	3	3	3

(a)

	0	1	2	3
0	2	2	1	0
1	0	4	0	0
2	0	0	5	2
3	0	0	0	4

(b)

Fig 3 a. Example of an image with 4 grey level image b. GLCM for distance 1 and direction 0°.

Figure 1.a. is an example matrix of pixels intensity representing image with 4 (four) levels of grey. Note the intensity level intensity level 0 and 1 are marked with a thin box. The thin box representing pixel-intensity 0 with pixel intensity 1 as its neighbour (in the horizontal direction or the direction of 0 °). There are two occurrences of such pixels. Therefore, the GLCM matrix formed (Fig. 1.b.) with value 2 in row 0, column 1. In the same way, GLCM matrix row-0 column 0 is also given a value of 2, because there are two occurrences in which pixels with value 0 has pixels 0 as its neighbour (horizontal direction). As a result, the pixels matrix representing in Figure 1.a. can be transformed into GLCM as Figure 1.b.

In addition to the horizontal direction (0 °), GLCM can also be formed for the direction of 45 °, 90 ° and 135 ° as shown in Figure 2 below.

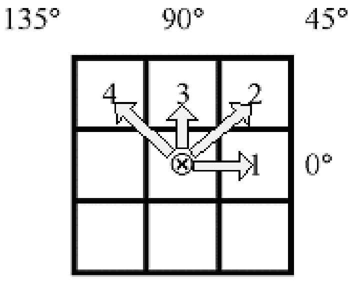


Fig 4 Direction of GLCM generation. From the center (⊗) to the pixel 1 representing direction = 0° with distance d =1, to the pixel 2 direction = 45° with distance d = 1, to the pixel 3 direction = 90° with distance d = 1, and to the pixel 4 direction = 135° with distance d = 1

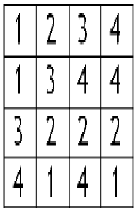
Haralick and his colleagues (1973: 613) extracting 14 features from the co-occurrence matrix, although in many applications only 8 (eight) features that are widely used, that is: Energy, Entropy, Max Probability, Inverse Diff. Moment, contrast, homogeneity, Inertia, and Correlation.

Although co-occurrence matrices capture the texture properties, it never directly used as a tool for analysis, such as comparing the two textures. The matrix of data must be extracted again to get the numbers that can be used to classify the texture. Haralick(1973) proposed 14 measures (or features), but Connors and Harlow in their study proposed, only 5 of 14 Haralick's features which are commonly used. These five features are: energy, entropy, correlation, homogeneity, and inertia (Kulak, 2002:62).

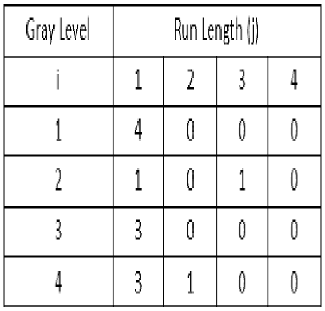
2.6. Gray-Level Run-Length Matrix

Grey-level run-length matrix (GLRLM) is a matrix from which the texture features can be extracted for texture analysis. Texture is understood as a pattern of grey intensity pixel in a particular direction from the reference pixels. Run length is the number of adjacent pixels that have the same grey intensity in a particular direction. Gray-level run-length matrix is a two-dimensional matrix where each element $p(i, j \mid \theta)$ is the number of elements j with the intensity i , in the direction □. For example, Figure 1 below shows a matrix of size 4x4 pixel image with 4 gray levels.

Figure 3 is a representation matrix GLRL (grey-level run-length) in the direction of 0° $[P(i,j\mid\theta=0^{\circ})]$



1	2	3	4
1	3	4	4
3	2	2	2
4	1	4	1



Gray Level	Run Length (j)			
i	1	2	3	4
1	4	0	0	0
2	1	0	1	0
3	3	0	0	0
4	3	1	0	0

Fig5MatrixofImage 4X4 pixels Fig 6 GLRL Matrix

In addition to the 0º direction, GLRL matrix can also be formed in the other direction, i.e. 45º, 90º or 135º

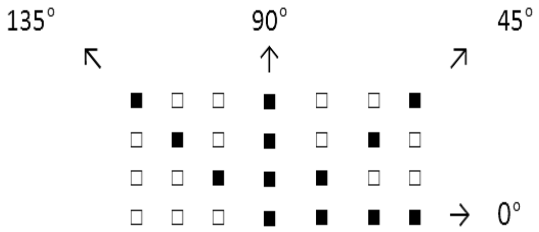


Fig 7 Run Direction

Some texture features can be extracted from the GLRL matrix. Galloway (Tang, 1998:1602-1609) suggests 5 texture features based on this GLRL matrix, namely: Shot Runs Emphasis (SRE), Long Runs Emphasis (LRE), Gray Level Non-uniformity (GLN), Run Length Non-uniformity (RLN), and Run Percentage (RP). Based on the observations that most of the features is only a function of $P_r(j)$, regardless of the grey level information contained in $P_g(i)$, Chu et al (1990:415-420) adds 2 more features called Low Gray Level Run Emphasis (LGRE) and High Gray Level Run Emphasis (HGRE). This feature uses grey level of pixels in sequence and is intended to distinguish the texture that has the same value of SRE and LRE but have differences in the distribution of gray levels. asarathy and Holder (Tang, 1998:1602-1609) added 4 more features extracted from the matrix GLRL, namely: Short Run Low Gray-Level Emphasis (SRLGE), Short Run High Gray Level Emphasis (SRHGE), Long Run Low Gray Level Emphasis (LRLGE), and Long Run High Gray Level Emphasis (LRHGE).

III. CLASSIFICATION

3.1 PREPARATION DATABASE

The extracted features are organized in a database in the form of transactions [27], which in turn constitute the input for deriving association rules. The transactions are of the form [Image ID, F1; F2; :::; F9] where F1:::F9 are 9features extracted for a given image.

3.2. ASSOTIATION RULE MINING

Discovering frequent item sets is the key process in association rule mining. In order to perform data mining association rule algorithm, numerical attributes should be discretized first, i.e. continuous attribute values should be divided into multiple segments. Traditional association rule algorithms adopt an iterative method to discovery, which requires very large calculations and a complicated transaction process. Because of this, a new association rule algorithm is proposed in this paper. This new algorithm adopts a Boolean vector method to discovering frequent item sets. In general, the new association rule algorithm consists of four phases as follows:

1. Transforming the transaction database into the Boolean matrix.

2. Generating the set of frequent 1-itemsets L1.

3. Pruning the Boolean matrix.

4. Generating the set of frequent k-item sets Lk(k>1).

The detailed algorithm, phase by phase, is presented below:

1. *Transforming the transaction database into the Boolean matrix:* The mined transaction database is *D*, with *D* having m transactions and *n* items. Let T={T1,T2,…,Tm} be the set of transactions and I={I1,I2,…,In}be the set of items. We set up a Boolean matrix Am*n, which has m rows and n columns. Scanning the transaction database *D*, we use a binning procedure to convert each real valued feature into a set of binary features. The 0 to 1 range for each feature is uniformly divided into k bins, and each of *k* binary features record whether the feature lies within corresponding range.

2. *Generating the set of frequent 1-itemset L1:* The Boolean matrix Am*n is scanned and support numbers of all items are computed. The support number Ij.supth of item Ij is the number of '1s' in the jth column of the Boolean matrix Am*n. If Ij.supth is smaller than the minimum support number, itemset {Ij} is not a frequent 1-itemset and the jth column of the Boolean matrix Am*n will be deleted from Am*n. Otherwise itemset {Ij} is the frequent 1-itemset and is added to the set of frequent 1-itemset L1. The sum of the element values of each row is recomputed, and the rows whose sum of element values is smaller than 2 are deleted from this matrix.

3. *Pruning the Boolean matrix:* Pruning the Boolean matrix means deleting some rows and columns from it. First, the column of the Boolean matrix is pruned according to Proposition 2. This is described in detail as: Let I• be the set of all items in the frequent set LK-1, where k>2. Compute all |LK-1(j)| where j belongs to I2, and delete the column of correspondence item j if |*LK* – 1(*j*)| is smaller than *k* – 1. Second, re-compute the sum of the element values in each row in the Boolean matrix. The rows of the Boolean matrix whose sum of element values is smaller than k are deleted from this matrix.

4. *Generating the set of frequent k-itemsets Lk:* Frequent k-item sets are discovered only by "and" relational calculus, which is carried out for the k-vectors combination. If the Boolean matrix *Ap*q* has q columns where 2 < *q* £ *n* and *minsupth* £ *p* £ *m*, *k q c,* combinations of k-vectors will be produced. The 'and' relational calculus is for each combination of k-vectors. If the sum of element values in the "and" calculation result is not smaller than the minimum support number *minsupth*, the k-itemsets corresponding to this combination of kvectors are the frequent k-itemsets and are added to the set of frequent k-itemsets Lk..

IV. RESULT AND DISCUSSION

A total of 23 malignant masses and of 65 benign masses images measuring 50x50 pixels is transformed into a grey level co-occurrence matrix (GLCM) and grey level run-length matrix (GLRLM). Based on GLCM, 8 texture features was extracted, and based on GLRLM 11 texture features was extracted. The result showed that the 19 texture features could differentiate malignant masses and benign masses with accuracy up to 94.9%, depending on the number of features using as predictors.

Table 1: Predicted Accuracy of Association rule mining for Texture Features Based onGLCM and GLRLM

GLCM_GLRL Texture Features	
All Features(19)	12 Features
94.9%	92.3%

Performance evaluation accuracy of statistical prediction model can also be done by ROC (receiver operating characteristics) curve analysis. ROC curve is a graphical plotting with the y-axis express sensitivity and the x-axis express false positive rate (Zou, et al, 2007: 654; Park, et al, 2004:11).The following figure shows the ROC curve for discrimination using features based on GLCM and GLRLM as the predictors for the image size 21x21 pixels.

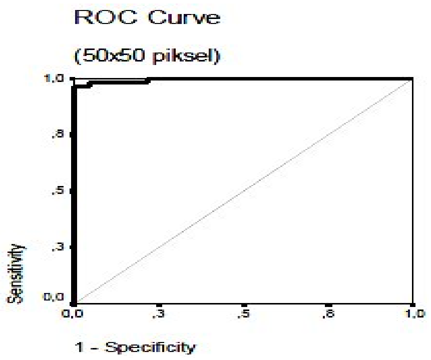
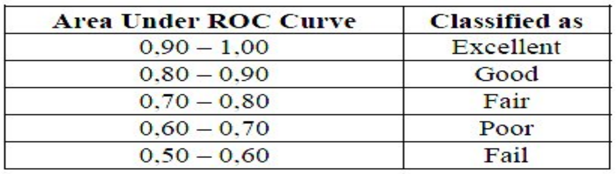


Fig8ROC Curve for Discrimination using GLCM and GLRLM Features

Figure 7 shows that its ROC curve is closer to the top y-axis, meaning that the level of prediction accuracy with texture features based on GLCM and GLRLM is 0.995, indicates that the prediction accuracy could be classified as very good.

Table 3: Classifying Level of Accuracy Based on Area Under ROC Curve



Area Under ROC Curve	Classified as
0.90 – 1.00	Excellent
0.80 – 0.90	Good
0.70 – 0.80	Fair
0.60 – 0.70	Poor
0.50 – 0.60	Fail

V. CONCLUSION

Based on research results and discussions, it concludes that:

1. Texture features based on GLRLM can be used to distinguish between malignant masses and benign masses on ultrasound images, with accuracy levels that are relatively lower than texture features based on GLCM and texture features based on combined GLRLM and GLCM.

2. Texture features based on GLCM can be used to distinguish between malignant masses and benign masses on mammogram images, with accuracy levels higher than texture features based on GLRLM, but still lower than texture features based on combined GLRLM and GLCM.

3. Important texture features to distinguish malignant masses and benign masses on mammograms are: SRE, LRE, GLN, RLN, LGRE, HGRE, SRLGE, Energy, Inertia, Entropy, Maxprob, Inverse.

VI. REFERENCES

[1] Majid AS, de Paredes ES, Doherty RD, Sharma N Salvador X. “Missed breast carcinoma: pitfalls and Pearls”. Radiographics 23(2003)881-895.

[2] Osmar R. Zaïane,M-L. Antonie, A. Coman “Mammography Classification by Association Rulebased Classifier,” MDM/KDD2002 International Workshop on Multimedia Data Mining with (ACM SIGKDD 2002, Edmonton, Alberta, Canada, 17-19 July 2002,), pp.62-69.

[3] Xie Xuanyang, Gong Yuchang, Wan Shouhong, Li Xi ,”Computer Aided Detection of SARS Based on Radiographs Data Mining “, Proceedings of the 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference Shanghai, China, September 1-4, 2005

[4] Shuyan Wang, Mingquan Zhou and Guohua Geng, “Application of Fuzzy Cluster analysis for Medical Image Data Mining” Proceedings of the IEEE International Conference on Mechatronics & Automation Niagara Falls, Canada • July 2005.

[5] R.Jensen, Qiang Shen, “Semantics Preserving Dimensionality Reduction: Rough and Fuzzy-Rough Based Approaches”, IEEE Transactions on Knowledge and Data Engineering, 16 (2004) 1457-1471.

[6] Walid Erray, and Hakim Hacid, “A New Cost Sensitive Decision Tree Method Application for Mammograms Classification” IJCSNS International Journal of Computer Science and Network Security, 6 (2006) No.11.

[7] Ying Liu, Dengsheng Zhang, Guojun Lu, “Region based image retrieval with high-level semantics using decision tree learning”, Pattern Recognition, 41 (2008) 2554 – 2570

[8] Kemal Polat , Salih Gu¨nes, A novel hybrid intelligent method based on C4.5 decision tree classifier and one-against-all approach for multi-class classification problems, Expert Systems with Applications,

[9] C.Chen and G.Lee, “Image segmentation using multitiresolution wavelet analysis and Expectation Maximum(EM) algorithm for mammography” , International Journal of Imaging System and Technology, 8(5):491-504,1997

[10] T.Wang and N.Karayaiannis, “Detection of microcalcification in digital mammograms using wavelets”, IEEE Trans. Medical Imaging, 17(4):498-509,1998

[11] I.Christiyanni et al ., “Fast detection of masses in computer aided mammography”, IEEE Signal processing Magazine, Pages:54- 64,2000.

[12] Deepa S. Deshpande “association rule mining based on image content” International Journal of Information Technology and Knowledge Management January-June 2011, Volume 4, No. 1, pp. 143-146

[13] L. Jaba Sheela & V.Shanthi “A Novel Texture Classification Procedure by using Association Rules “ITB J. ICT Vol. 2, No. 2, 2008,pp,103-114

[14] http://marathon.csee.usf.edu/Mammography/Database.html

[15] Ali Cherif chaabani, Atef boujelben, Adel mahfoudhI, Mohamed ABID, "An Automatic-Pre-processing Method For Mammographic Images ", JDCTA: International Journal of Digital Content Technology and its Applications, Vol. 4, No. 3, pp. 190 ~ 201, 2010

[16] M. Hanmandlu, V.K. Madasu and S. Vasikarla, “A fuzzy Approach to texture segmentation”, Proc. International Conference on Information Technology: Coding and Computing, vol-1, pp.636-642, 2004.

[17] V. Chalana and Y. Kim, “A Methodology for Evaluation of Boundary algorithms on Medical Images,” IEEE Trans. on Medical Imaging, vol. 16(5), pp. 642-52, 1997.

[18] W. A. Perkins, “Area segmentation of images using edge points.” IEEE Trans. on Pattern Analysis & Machine Intelligence, vol. PAMI-2, pp. 8-15. 1980.

[19] G. Torheim, F. Godtliebsen, D. Axelson, K. A. Kvistad, O. Haraldseth, and P. A. Rinck, "Feature extraction and classification of dynamic contrast-enhanced T2*-weighted breast image data," IEEE Transactions on Medical Imaging, vol. 20, pp. 1293-301, 2001.

[20] Abu Sayeed Md. Sohail, Prabir Bhattacharya, Sudhir P. Mudur and Srinivasan Krishnamurthy, “Classification of Ultrasound Medical Images Using Distance Based Feature Selection and Fuzzy-SVM”,Pattern Recognition and Image Analysis,Lecture Notes in Computer Science, 2011, Volume 6669/2011, 176-183, DOI: 10.1007/978-3-642-21257-4_22

[21] R. M. Haralick, K. Shanmugam, and I. Dinstein, “Textural features for image classification”, IEEE Trans. System Man. Cybernetics, vol. SMC-3, pp. 610–621, 1973.

[22] R. Gupta and P.E. Undrill, “The use of texture analysis to identify suspicious masses inmammography”, Phys. Med. Bio., vol 15. 835- 855, 1997.

[23] Chan et al., “Computer-aided classification of mammographic masses and normal tissue: linear discriminant analysis in texture feature space”, Phys. Med. Biol., vol. 40, 857-876.

[24] D. Wei, H. P. Chan, M. A. Helvie, B. Sahiner, N. Petrick, D. D. Adler, and M. M. Goodsitt,"Classification of mass and normal breast

tissue on digital mammograms: multiresolution texture analysis," Medical Physics., vol. 22, pp. 1501-13, 1995.

[25] [25]P. Gibbs and L. W. Turnbull, "Textural analysis of contrast-enhanced MR images of the breast,"Magnetic Resonance in Medicine, vol. 50, pp. 92-8, 2003.