A Novel Approach for Automatic detection of cancerous masses in mammogram MRI

[(1)]S.Pitchumani Angayarkanni, [(2)]Dr.Nadira Banu Kamal,[(3)] Mrs.V.Thavavel
Department of Computer Science, Lady Doak College,Madurai, Tamil Nadu, India
Department of M.C.A, TBAK College ,Kilakarai,Ramnad,Tamil Nadu,India
Department of M.C.A., Karunya University,Coimbatore,TamilNadu,India

Abstract

Breast cancer is one of the most common forms of cancer in women. In order to reduce the death rate, early detection of cancerous regions in mammogram images is needed. The existing system is not so accurate and also time consuming. The proposed system is mainly used for automatic segmentation of the mammogram images to classify them as benign, malignant or normal based on the decision tree ID3 algorithm. A hybrid method of data mining technique is used to predict the texture features which play a vital role in classification. Automatic classification is done through 3 stages ANN. The weights in ANN are adjusted using the rule derived from ID3 algorithm. The sensitivity, the specificity, positive prediction value and negative prediction value of the proposed algorithm accounts to 9.78%, 99.9%, 94% and 98.5% which rates very high when compared to the existing algorithms. This paper focuses on the comparative analysis of the existing methods and the proposed technique in terms of sensitivity, specificity, accuracy, time consumption and ROC.

Keywords: GLCM,Gabor filter, SOM,ANN and Data mining techniques.

1. Introduction

Breast cancer has been determined to be the second leading cause of cancer death in women, and the most common type of cancer in women. The mammography is the best method of diagnosis by images that exists at the present time to detect minimum mammary injuries, fundamentally small carcinomas that are shown by micro calcifications or tumors smaller than 1cm. of diameter that are not palpated during medical examination. [Antonie *et al*, 2001]. Currently, joint efforts are being made in order to detect tissue anomalies in a timely fashion, given that there are no methods for breast cancer prevention. Early detection has proved an essential weapon in cancer detection, since it helps to prolong patients' lives. Physicians providing test results must have diagnostic training based on mammography, and must issue a certain number of reports annually. Double reading of reports increases sensitivity for detection of minimal lesions by about 7%, though at a high cost. The physician shall then interpret these reports and determine the steps to be taken for the proper diagnosis and treatment of the patient. For this reason, physicists, engineers, and physicians are in search of new tools to fight cancer, which would also allow physicians to obtain a second opinion [Gokhale *et al*, 2003, Simoff *et al*, 2002]. Different methods have been used to classify and/or detect anomalies in medical images, such as wavelets, fractal theory, statistical methods and most of them used features extracted using image processing techniques [1]. In addition, some other methods were presented in the literature based on fuzzy set theory, Markov models and neural networks.

Most of the computer-aided methods proved to be powerful tools that could assist medical staff in hospitals and lead to better results in diagnosing a patient [Antonie *et al*, 2001]. Different studies on using data mining in the processing of medical images have rendered very good results using neural networks for classification and grouping. In recent years different computerized systems have been developed to support diagnostic work of radiologists in mammography[2].

The proposed method includes the following phases i)Image Pre-processing and enhancement ii)Segmentation iii)Classification using ID3 Algorithm, iv) Predicting size and stages v) ANN and vi) Accuracy of algorithm prediction.

PSNR	RMS	NSD	ENL	MES	Nature of Filter
87.65	2.97	4.55	89.89	8.83	Gabor

Table 1: Signal to Noise ratio calculation

2. Image Pre-processing and Enhancement

The main objective of pre-processing is to enhance the image and remove unwanted data. This is done by using gabor filter and histogram equalization. Gabor wavelet filters smooth the image by blocking detailed information. Mass detection aims to extract the edge of the tumor from surrounding normal tissues and background[1]. PSNR, RMS, MSE, NSD, ENL value calculated for each of 121 pairs of mammogram images clearly shows that gabor wavelet filter when applied to mammogram image leads to best Image Quality[4]. The orientation and scale can be changed in this program to extract texture information. Here 3 scales and 4 orientation was used[9].

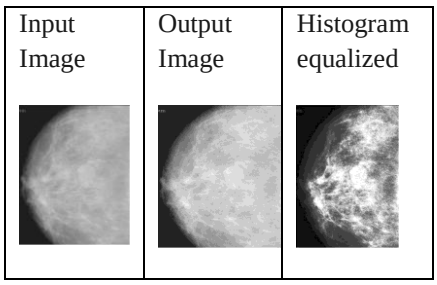


Figure 1: Results after preprocessing and enhancement

The results clearly show that the gabor filter with histogram equalization produces high PSNR value indicating that the image is highly enhanced.It also removes tape artifact, high intensity rectangular label and low intensity labels.

2.1Comparison of Existing with proposed technique:

Filters Used	Features & Limitations	PSNR value
Median Filter	Used to smooth the non repulsive noise from two-dimensional signals without blurring edges and preserve image details. Suitable for enhancing mammogram images. Pectoral muscles are not detected	72.54
LowPass Filter	Reduces noise and also blurs the edges	65.45
Highpass filter	Enhances the details of the image	68.79
Partial low and high pass filters	Best Quality image is achieved	84.90
Spatial and frequency domain filters	Used for image enhancement alone	55.67
Proposed Method	**Features**	**PSNR**
Gabor Filter with histogram equalization	• It acts as a local band-pass filter with optimal number of orientations and the scales define the number of filters that should affect input images by multiplying them with each other. • The joint localization properties of the image is enhanced by histogram equalization in the spatial and in the spatial frequency domain. It is used for detecting a first set of potential microcalcifications and elongated structures are identified • It also detects clusters of microcalcifications to extract textural features of an image	89.97

3.Texture based Approach and SOM based Visualization:

Texture based segmentation is implemented because for a person affected by cancer the texture of the skin becomes smooth. This segmentation method segments the calcification pattern and the other suspicious regions in the mammograms The GLCM image is divided into 3x3 matrix and the texture features are calculated[2,3].

Texture features are: Cluster prominence,Energy,Entropy,Homogenity,Difference variance, Difference Entropy, Information Measure, Normalized ,Correlation.

Using GLCM (Gray Level Co-Occurrence Matrix) technique, the different combinations of brightness values that occur on the texture segmented image is found. Usually the GLCM matrix is found for small windows but in this project the GLCM matrix is found for the whole image. Then the GLCM Matrix is divided into small windows of size 3x3. Since the size of the Mammogram is larger, the size of the image is resized to 17x17 and the GLCM matrix gets segmented into 289 images. GLCM features: Cor-relation, Cluster Prominence, Energy, Entropy, Homogeneity, Difference Variance, Difference Entropy and Information Measure related to Cor-relation, and Normalized are calculated and stored in an Excel file.The texture values for 121 pairs of Mammogram MRI images are calculated and are stored in an excel sheet and it is analysed using SOM based Visualization technique .

Pseudo code for performing Texture segmentation is:

- Step 1: Read Image
- Step 2: Create Texture Image
- Step 3:Create Rough Mask for the background Texture
- Step 4: Use Rough Mask to Segment the Foreground Texture
- Step 5: Display Segmentation Results

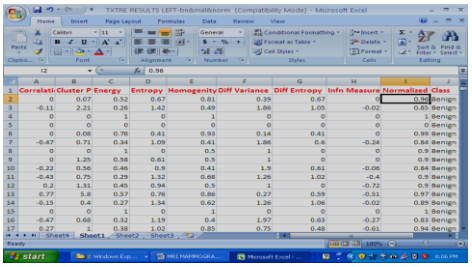
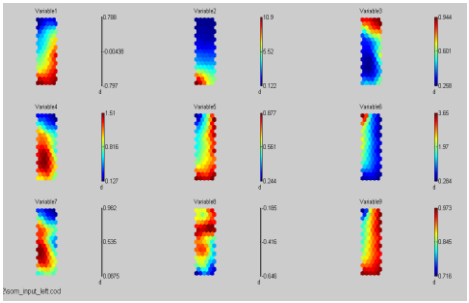
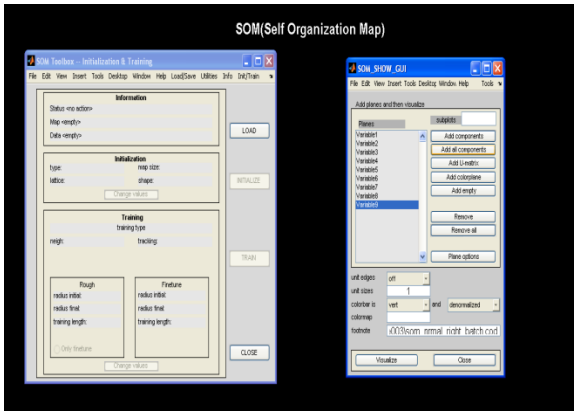


Table 2: The Texture Parameter value for the left Mammogram Image – Benign Case

The unified distance matrix or U-matrix is a representation of the Self-Organizing Map that visualizes the distance between the network neurons or units. It contains the distance from each unit center to all of its neighbors. The neurons of the SOM network are represented here by hexagonal cells. The distance between the adjacent neurons is calculated and presented with different colorings. A dark coloring between the neurons corresponds to a large distance and thus represents a gap between the values in the input space. A light coloring between the neurons signifies that the vectors are close to each other in the input space. Light areas represent clusters and dark areas cluster separators. This representation can be used to visualize the structure of the input space and to get an impression of otherwise invisible structures in a multidimensional data space. The U-matrix representation (Figure 2) reveals the clustering structure of the dataset explored(Texture parameter) in this experiment. Texture parameters having similar characteristics are arranged close to each other and the distance between them represents the degree of similarity or dissimilarity.



A) Visualization Output



B) SOM TOOLBOX

Figure 2: SOM based visualization for Benign case using SOM Toolbox

The SOM toolbox produces secondary, strengthened, features which can then be used to segment or classify the image according to the texture energy. SOM toolbox in this research has helped to visualize the relationship between the features and also how the feature varies for different types of cases like Benign, Malignant and Normal.The output clearly indicates that 'Information Measure related to Correlation' varies for the above specified cases during Mapping and it is also found that Energy and Entropy are oppositely correlated.

4. Decision Tree Induction Method, ID3 Algorithm:

- A mathematical algorithm for building the decision tree.
- Invented by J. Ross Quinlan in 1979.
- Uses Information Theory invented by Shannon in 1948.
- Builds the tree from top down, with no backtracking.

Information Gain is used to select the most useful attribute for classification.

Entropy

- A formula to calculate the homogeneity of a sample.
- A completely homogeneous sample has entropy of 0.
- An equally divided sample has entropy of 1.
- Entropy(s) = - p+log2 (p+) -p-log2 (p-) for a sample of negative and positive elements.
- The formula for entropy is:

$$Entropy(S) = \sum_{i=1}^{C} p_i \log_2 p_i$$

Information Gain (IG):

- The information Gain is based on the decrease in entropy after a dataset is split on an attribute[7].

Information Gain(S,A)=Entropy(S)-H(S,A)

Where $H(S,A)=\sum_i(|Si|/|S|).H(Si)$

A takes on value 1 and **H(Si)** is the entropy of the system of subsets Si.

The training data is a set S=s1,s2—of already classified samples based on the texture features. Each sample Si=x1,x2 is a vector where x1,x2—represents attributes or features of the sample. The training data is augmented with a vector C=C1,C2 and C3 where C1 represents benign, C2 represents malignant and C3 represents normal cases. At each node of the tree ID3 chooses one attribute of the data that most efficiently splits its set of samples into subsets enriched in one class or the other. The criterion is the normalized Information Gain that results from choosing an attribute for splitting the data. The attribute with highest information Gain is chosen to make decision.

Pseudo code:

Input: Set of texture feature attributes A1,A2---An. The class labels Ci.The number of classes Pi. Training set S of examples

Output: Decision tree with set of rules.

Procedure:

- All the samples in the list belongs to three different classes
- Create a leaf node for the decision tree to choose the class.
- If none of the feature provides IG then ID3 creates a decision node higher up the tree using the expected value of the class.
- If instance previously unseen class is encountered, ID3 creates decision class higher up in the tree using expected value.
- Calculate IG for each attribute
- Choose attribute A with lowest entropy and Highest IG and test this attribute with root.
- For each possible value v of this attribute
    - Add a new branch below the root corresponding to A=v.
    - If v is empty make the new branch a leaf node labeled with most common value else
    - Let the new branch be the tree created by ID3

- End

The attributes used were the nine texture parameters with the class as benign, malignant and normal . Based on the rule derived by testing 121 pairs of various mammogram images the rules are applied in classifying the new cases without prior knowledge of whether they were benign, malignant or normal[8].

An example of ID3 decision tree classification applied for a left benign case of mammogram MIAS dataset.

Node1	Cluster Prominence
Node2	Energy
Node3	Entropy
Node4	Homogenity
Node5	Difference variance,
Node6	Difference Entropy,
Node7	Information Measure,
Node8	Normalized
Node9	correlation
Class	Benign/Malignant/Normal

Table 3: Nodes representing the 10 attributes

Cross validation 10 fold

=== Detailed Accuracy By Class ===

Class	TP Rate	FP Rate	Precision	Recall	FMS	ROC
Benign	0.595	0.272	0.523	0.595	0.557	0.677
Malig	0.121	0.104	0.368	0.121	0.182	0.542
Normal	0.678	0.427	0.442	0.678	0.536	0.637
Wgt.	0.465	0.268	0.445	0.465	0.425	0.618

=== Confusion Matrix ===

a b c <-- classified as

172 30 87 | a = Benign

94 35 160 | b = Malignant

63 30 196 | c = Normal

Error rate	0.5409

Values prediction

Value	Recall	1-Precision
Benign	0.6090	0.4854
Malignant	0.0000	1.0000
Normal	0.7682	0.5771

Confusion matrix

	Benign	Malignant	Normal	Sum
Benign	176	0	113	28
Malignant	99	0	190	28
Normal	67	0	222	28
Sum	342	0	525	86

The classification accuracy obtained from these fixed-order trees can be compared with those from trees of different feature orders, **as** well **as** with those from trees of different feature combinations. The decision tree with the optimal feature combination and order for this task can thus be identified. Examples of the training and test results obtained in this study will be discussed below. It should be noted that, to use a trained decision-tree classifier, one has to choose the specific tree structure with the set of decision thresholds corresponding to the desired sensitivity (TPF) and specificity (~-FPF)**.**The structure and the thresholds will be fixed during testing or application. The decision rule is indicated below

- Infn Measure < -0.0450 then Class = **Benign** (64.89 % of 262 examples)
- Infn Measure >= -0.0450 then Class = **Malignant** (62.34 % of 316 examples).
- Infn Measure >= -0.0050 then Class = **Normal** (42.29 % of 525 examples)

ID3 parameters:
Size before split 200
Size after split 50
Max depth of leaves10
Goodness of split threshold 0.0300

Classifier performances

This rule is then applied for classifying the new dataset of segmented mammogram images as benign , malignant and normal cases automatically. The results of the prediction were checked with the clinically proven classification results by the radiologist and were found that the rule derived provides 100% accuracy in classification.

S. No	No. of cases trained	No. of cases tested	Normal Trained/ tested	Benign Trained/ tested	Malignant Trained/tested
1	40 Pairs	20 Pairs	3/6	25/4	22/10

Table 4: No. of cases tested and trained

Statistical Method for stage prediction

Cancer **stage** is based on four characteristics:

- the size of the cancer
- whether the cancer is invasive or non-invasive
- whether cancer is in the lymph nodes
- whether the cancer has spread to other parts of the body beyond the breast.

Table 1: Predicting Cancer stages

The size of the cancer is detected using the ellipsoid volume formula for the ROI using the formula

π/6 x L x W x H and 1/2 x L x W x H.

Where L-length,W-Width and H-Height .

Of the entire sample of women diagnosed with invasive breast cancer, 51% had stage I; 26% stage II; 11% stage III; and 4% stage IV disease. The equation predicting stage IV disease achieved sensitivity of 81%, specificity 89%, positive predictive value (PPV) 24%, and negative predictive value (NPV) 99%, while the equation distinguishing stage I/II from stage III disease achieved sensitivity 83%, specificity 78%, PPV 98%, and NPV 31%. The equations most accurately identified early stage disease and ascertained a sample in which 98% of patients were stage I or II.

No. of Cases detected	Size	Lymph node	Stage
51	<=2cm	Not Detected	I
26	3 cm	Partially Detected	II
11	6 cm	Fully Detected	III
4	>6cm	Spread found other than lymph node	IV

Table 5: Predicting Cancer stages

5. NEURAL NETWORK CLASSIFICATION

Multilayered feed forward neural network is used for training in this proposed method. Reason for choosing multilayer BPN is that it involves non parametric statistical properties. Unlike the classical statistical classification methods, such as bayes classifier ,no knowledge of the underlying probability distribution is needed by a neural network. It can learn the free parameters (weight and bias) through training by example. This makes it suitable to deal with real problems which are nonlinear, non stationery and non Gaussian. The neural network classifier is used to generate a likelihood map of each mammogram using gabor feature as input to classifier.

A neural network for solving classification problem typically has N input neurons and M output neurons. The k^{th} output neuron (1<=K<=M) is trained to output one for pattern belonging to the k^{th} class. A single output neuron suffices in the case of a problem that involves two categories of classification.ID3 based multilayered perceptron facilitates the classification of non linear problems .It is simpler to isolate the problem if the number of hidden neurons are high. The proposed method for the system is learning by trial and error which consists of adjusting weights of the connections according to the IG parameter value derived from ID3 algorithm with $\pm Eq=1/2\sum(r_Q-o_Q)^2$

Where r_Q is Quadratic error among expected response and o_Q is current response.

The feature set extracted from 285 gabor feature set classified using the ID3 based Decision tree induction is used to classify benign, malignant and normal cases .These feature sets are given as input to the network for training. The desired output from the network is whether the classification is malignant, benign or normal tissue. Four neurons are used in hidden layer and the nine feature sets are used in input layer. During the training session of the network a pair of pattern is presented, the

input pattern and target pattern (Malignant,Benign and Normal). At the output layer, the difference between the actual and target output yields an error signal. This error signal depends on the values of the weights of the neurons in each layer. This error is minimized and during these processes new values for the weights are obtained.

Data partition results:

Data partition results:

575 records to Training set (68.05%)

135 records to Validation set (15.98%)

135 records to Test set (15.98%)

Data anomalies:

22 numeric outliers

Network architecture: [9-4-3]

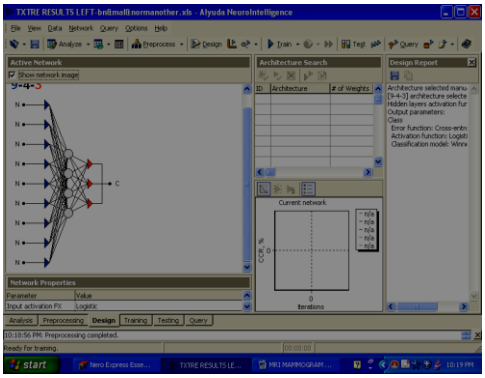
Training algorithm: Online Back Propagation

Number of iterations: 601

Time passed: 00:00:06

Training stop reason: All iterations done

The best network was tracked and restored



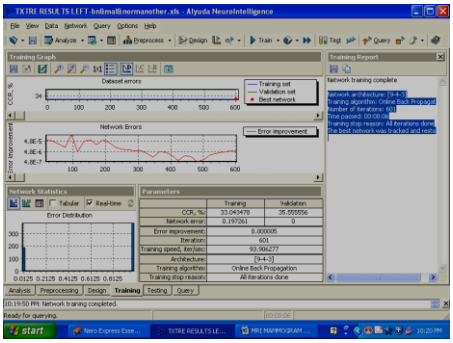


Figure 3: ANN Error Distribution

Methods	**Author and References**	**Computational Time**
Morphological Analysis	Wan Mimi Diyana, Julie Larcher, Rosli Besar	3’20’’
Filtering Technique	Proposed Approach	3’50’’
Fractal Dimension Analysis	Wan Mimi Diyana, Julie Larcher, Rosli Besar	7’20’’
Complete HOS Test	Wan Mimi Diyana, Julie Larcher, Rosli Besar	9’20’’
neuro-fuzzy segmentation	S. Murugavalli et. al	93’ 39’’
Proposed Method	S.Pitchumani Angayarkanni ,Nadhira Banu Kamal	0’03”

Table 6: Comparative analysis of Computational time.

6. ROC Analysis:

ROC analysis is based on statistical decision theory, developed in the context of electronic signal detection, and has been applied extensively to diagnostic systems in clinical medicine. The ROC curve is a plot of the classifier’s true positive detection rate versus its false positive rate. The false positive (FP) rate is the probability of incorrectly classifying a non-target object (e.g. normal tissue region) as a target object (e.g. tumor region). Similarly,

the true positive (TP) detection rate is the probability of correctly classifying a target object as being a target object. The TP and FP rates both are specified in the interval from 0.0 to 1.0, inclusive, in medical imaging, the TP rate is referred to as sensitivity, and (1.0 – FP rate) is called specificity. Statistical classifiers have parameters that can be varied to alter the TP and FP rates. Using these parameters, an ROC curve can be generated which shows the TP/FP trade-off associated with the different values that the parameter(s) may assume. It would then be possible to trade a lower (higher) FP rate for a higher (lower) TP detection rate by choosing the appropriate value(s) for the parameter(s) in question.

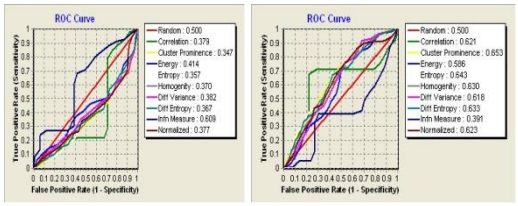


Figure 5: ROC curve for Benign and Malignant case

Partition	Sensitivity	Specificity	Accuracy
Training	99.58%	99.80%	99.99%
Test	100.00%	100.00%	100.00%

Table 7: Accuracy of the proposed algorithm

8. Conclusion:

The rule generated using decision tree induction method clearly shows that the time taken to classify benign malignant and normal cases in just 0.03 seconds and the accuracy is 99.9%.

The specificity =t-neg/neg, sensitivity =t-pos/pos and Precision=t-pos/t-pos+f-pos and accuracy=sensitivity pos/(pos+neg))+specificity(neg/(pos +neg)). (Positive Prediction Value) PPV: True positive / (true positives + false positives) PPV: 94% and (Negative Prediction value) NPV: True negative / (true negative + false negatives)=98.5%.From the above algorithm, the accuracy was found to be 99.9%. The proposed method yields a high level of accuracy in a minimum period of time that shows the efficiency of the algorithm.

So far the weights in ANN are not adjusted using decision tree rule evolved through data mining. This new concept has reduced the error rate and increased the efficiency.

Author & References	Methods	Detection Rate
Ferrari & Rangayyan	**Directional filtering with Gobar wavelets**	**74.4%**
Lau and Bischof	**Asymmetry Measures**	**80.0%**
Sallam and Bowyer	**Unwrapping Technique**	**81.6%**
AbuBaker, A. Qahwaji, R. ; Ipson, S.	first and second order statistical texture analysis techniques.	98%
Du-Yih Tsai,Yongbum Lee, Masaru Sekiya,Masaki Ohkubo	Gaussiandistributed fuzzy membership functions	88.5%

Fatima Eddaoudi , Fakhita Regragui	Texture based on Haralick features with SVM classification purpose	88.5%
T.J. Jose and P. Mythili	Content Based Image Retrieval	97.6%
Anamika Ahirwar 1, R.S. Jadon	*SOM and Fuzzy c-means clustering*	Not specified
Maria-Luiza Antonie et. al	neural networks using back-propagation with association rule mining	97.8%
S.SAHEB BASHA et. al	segmentation and fuzzy c- means	82%
G M Harshavardhan ; K Pranaw ; P Deepa Shenoy ; K R Venugopal ; L M Patnaik	random forest decision classifier	90%
Duckwon Chung ; Revathy, K. ; Eunmi Choi ; Dugki Min	fractal dimension and fractal signature	98%
Nahla Ibraheem Jabbar and Monica Mehrotra	fuzzy kohonen neural network	DETECTING ONLY MALIGNANT CASE
Shu-Ting Luo & Bor-Wen Cheng	decision tree (DT), support vector machine—sequential minimal optimization	82.2%
V.Sivakrithika, B.Shanth	Neural Network classifier	73.6%
S.Pitchumani ,Nadira	ID3 based ANN	99.9%

Table 8: Comparison of existing and proposed methods

References:

1. Randy P. Broussard et al, Gabor filtering for improved microcalcification detection in digital mammograms, *British Journal of Radiology (2007) 80*

2. Arnau Oliver, Jordi Freixenet, Joan Martí, Elsa Pérez,et al., A review of automatic mass detection and segmentation in mammographic images, **Medical Image Analysis Vol. 14, Issue 2, Pages 87-110**

3. Elodia B. Cole, Etta D. Pisano, Donglin Zeng, Keith Muller,et al. he Effects of Gray Scale Image Processing on Digital Mammography Interpretation Performance, **Academic Radiology Vol. 12, Issue 5, Pages 585-595**

4. Du-Yih Tsai,Yongbum Lee,Masaru Sekiya,Masaki Ohkubo, Medical image classification using geneticalgorithm based fuzzy-logic approach, Journal of Electronic Imaging 13(4), 780–788 (October 2004)

5. T.J. Jose and P. Mythili, 2009. Neural Network and Genetic Algorithm Based Hybrid Model for Content Based Mammogram Image Retrieval. *Journal of Applied Sciences, 9: 3531-3538*

6. Maria-Luiza Antonie **,** Osmar R. Za¨ıane **,** Alexandru Coman,**Application of Data Mining Techniques for Medical Image Classification,** International Journal of Computer Science & Information Technology (IJCSIT), Vol 3, No 1, Feb 2011

7. Aswini Kumar Mohanty, Saroj Kumar Lenka, **Efficient Image Mining Technique for Classification of Mammograms to Detect Breast Cancer,** Special Issue of IJCCT Vol. 2 Issue 2, 3, 4; 2010 for International Conference [ICCT-2010], 3rd-5th December 2010

8. **Aswini kumar mohanty, Sukanta kumar swain,Pratap kumar champati ,Saroj kumar lenka** M, Image Mining for Mammogram Classification by Association Rule Using Statistical and GLCM features, IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 5, No 3, September 2011

9. Harald Fischer* and Ju¨ rgen Hennig, Neural Network-Based Analysis of MR Time Series Magnetic Resonance in Medicine 41:124–131

10. L. Vibha , G M Harshavardhan ; K Pranaw ; P Deepa Shenoy ; K R Venugopal ; L M Patnaik , Classification of Mammograms Using Decision Trees, Database Engineering and Applications Symposium, 2006. IDEAS '06. 10th International,

pg263-266

Biography:


Iam an Assiatnt Professor in Computer Science at Lady Doak College.I received my B.Sc degree in Spl. Physics at Lady Doak College , M.C.A in Sri Meenakshi Govt. College for women and M.Phil in Madurai Kamaraj University.Currently Iam persuing my Ph.D at karunya University,Coimbatore.My areas of interest are Image Processing and Data Mining.


I am a Professor of the Department of Computer Science of TBAK College for Women, Kilakarai, Ramanathaputam District, Tamil Nadu, India. I did my B Sc in Madras University, MSc in Anna University, MPhil in Alagappa University and PhD in Manonmanium Sundaranar University. My area of Interests and research are Digital Image Processing and Software Engineering. I am at present supervising 5 research scholars for their PhD degree and have guided many MPhil scholars


Dr.Thavavel received M.C.A and M.Phil Degrees from Madurai Kamaraj University, India. She is an Assistant Professor(SG) in M.C.A at Karunya University Coimbatore, India. Her research area of interest involves medical image reconstruction in an Object Oriented approach and application of Genetic algorithms to medical image analysis.