ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 8, August 2013

Detection of masses in mammogram images

Heena Shaikh[1], D.A.Kulkarni[2],G.R.Udupi[3]

CSE Department, KLS Gogte Institute of Technology, Belgaum, India[1,2]

Department of E & C, KLS VDRIT Haliyal, India[3]

Abstract: Breast cancer is the most common cancer in women and is the second leading cause of cancer death. Although it is curable when detected early, about one third of women with breast cancer die, so it is one of the most dangerous types of cancer caused all over the world. In the last decade, many research projects have been carried out aiming to develop computational systems to help specialists in the task of interpreting radiological images. Therefore detection of masses in mammogram images can be used for the early detection of breast cancer. Main contributions of this study are demonstrating the potential of cellular neural networks(CNN) to segment suspect regions in mammographic images and proposing a methodology that includes use of geostatistical functions (Ripley's K function and Moran's and Geary's indices) as texture signatures for mass detection. In the first stage of methodology the image is acquired from the DDSM database which is then pre-processed and later segmented using CNN, further the feature extraction process is carried using geostatistical functions which are later classified using support vector machine. The proposed work can allow this methodology to be added as a computer tool for the medical area, providing support to specialists especially in cases in which visualization is difficult. This allows optimizing the features for higher efficiency.

Keywords: Mammogram, Cellular Neural Networks (CNN), geostatistical functions, Probabilistic Neural Networks.

I. INTRODUCTION

Cancer is a group of diseases that cause cells in the body to change and grow out of control. Most type of cancer cells finally form a lump or mass called a tumor, and hence it is named after the part of the body where the tumor originates. Breast cancer occurs with a high frequency among the world's population [1]. Breast cancer begins in breast tissue, which comprises of glands for milk production called lobules, and the ducts connect the lobules to the nipple. The remainder of the breast is composed of fatty, connective and lymphatic tissue as shown in figure 1[2].

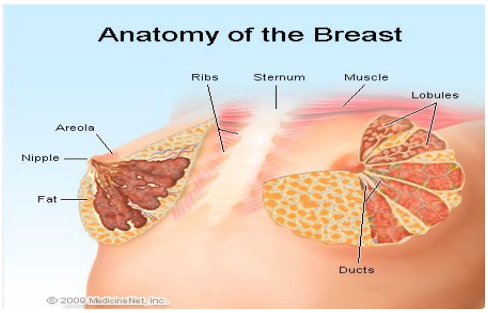

Figure 1: Anatomy of the Breast

The frequency of breast cancer in women has increased worldwide. It happens to over 11% of women during their life time. The world health organization's International Agency for Research on Cancer (IARC) estimates that more than a million cases of breast cancer will occur worldwide annually and more than 400,000 women die each year from this disease [1]. The X-ray mammography is a most effective and reliable method to detect the masses in early diagnoses and also helps in reducing high death rate caused due to breast cancer. Digital mammography is a convenient and easy tool in detecting and classifying them into masses and non-masses.

Statistics of breast cancer incidence and mortality in India is shown in following table 1 [3]:

Age-standardized rates (per 1,00,000 women)	India	Less developed regions	More developed regions
Incidence (%)	22.9	27.3	66.4
Mortality (%)	11.1	10.8	15.3

Table 1.1: Estimated breast cancer incidence and mortality in India by Age.

In 2008, 22.9% age-standardized incident cases were diagnosed per 100,000 women where as in 2008, 11.1% age-standardized breast cancer deaths were estimated per 100,000 women. In 2010 the total numbers of deaths estimated in India were 90,659. As against an estimated 48,170 women who died of breast cancer in the year 2007,

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 8, August 2013

the number breached 50,000 in the year 2010. The figure for the year was 50,821.

This work is organized as follows. In Section 2 the materials and methods required for the project work are presented. Section 3 represents the system design and the stages it involves such as acquisition, pre-processing the image, segmentation, feature-extraction and classification. Section 4 presents the techniques for the feature extraction and classification. Next, in Section 5, the results are shown in terms of figures. Finally, Section 6 presents some concluding remarks.

II MATERIALS AND METHODS

A. *Support Vector Machine (SVM)*

Support Vector Machine [4, 5] is a supervised learning model which is associated with learning algorithms that analyze data and recognize patterns that is used for classification and regression analysis.

The data set is divided into two sets: training and test. First, the sample is separated in two groups: masses and non-masses. Next, each group is randomly divided into10 subsets, from which one subset is chosen for training and the remaining ones are used for test. This process is repeated until all subsets have been tested. The support vector machine (SVM) was used with radial kernel and standard parameters (C=1 and y=0.5). A classification task usually involves separating data into training and testing sets. The goal of SVM is to produce a model (based on the training data) which predicts the target values of the test data given only the test data attributes. Since the proportion of non-masses selected in the segmentation stage is approximately six times higher than the number of masses, higher weight was assigned to the training of masses. This means that in training, the penalty for a mass classification error is greater than for a non-mass. A good balance between these two indices was achieved using weight 9 for the mass sample and weight 1 for the non-mass sample.The classification performance was analyzed by using the complete set of features and by applying step wise selection method for reducing and selection of features.

B. *Probabilistic Neural Networks (PNN)*

Probabilistic neural network [6] is a feed forward neural network derived from the Bayesian network. In a PNN, the operations are organized into a multilayered feed forward network. It consists of four layers which are as follows:

- Input layer
- Hidden layer
- Pattern layer
- Output layer

PNN is useful neural network architecture with slight difference in fundamentals from back propagation. The architecture is feed forward in nature which is similar to back propagation, but learning occurs in a different way. PNN is a supervised learning algorithm but includes no weights in its hidden layer. Each hidden node represents an example vector such that example acting as the weights to that hidden node. Figure 2 [8] illustrates a sample PNN.

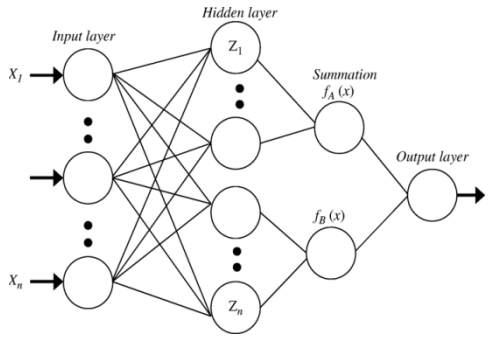

Figure 2: PNN architecture

PNN consists of an input layer, which represents the input pattern or feature vector. The input layer is interconnected with the hidden layer, which includes example vectors i.e. the training set for the PNN. The real example vector serves as the weights which are applied to the input layer. Lastly, an output layer represents each of the possible classes for which the input data can be classified. Nonetheless, the hidden layer is not fully interconnected to the output layer. The example nodes for a given class are connected only to that class's output node and none other.

One more important element of the PNN is the output layer and also the determination of the class for which the input layer fits in properly. This is done using a winner-takes-all approach. The output class node with the largest activation corresponds to the winning class. The class nodes are connected only to the example hidden nodes for their class where as the input feature vector connects to all examples, and thus influences their activations. It is therefore the sum of the example vector activations that determines the class of the input feature vector.

III SYSTEM DESIGN

The proposed methodology aims to develop a system to detect masses in mammogram image in an early stage of breast cancer. The system comprises of the following stages: acquisition, pre-processing, segmentation of regions of interest, feature extraction and classification [7] of the regions of interest as mass or non-mass. Figure 3 represents system design illustrating the stages.

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 8, August 2013

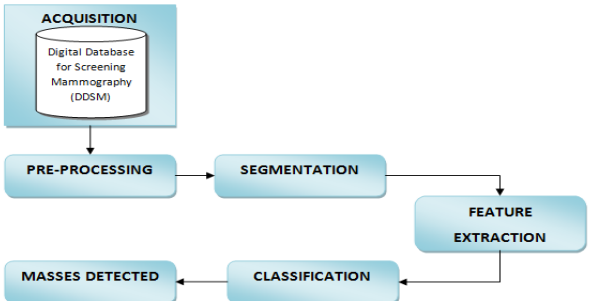

Figure 3: System Design

The System Design is divided into two sections; Training and Testing. In training, we train some data set of mammogram images in order to get correct test results. Training session contains the pre-processing, segmentation and feature-extraction stages, while the testing session consists of classification stage, where the classification is done by two models SVM and PNN, where the classification the candidate regions can be done into mass or non-mass. Finally, these methods are compared in terms of specificity and sensitivity.

IV. IMPLEMENTATION

The implementation details are discussed below.

1. *Acquisition*

Digital Database for Screening Mammography (DDSM) [9, 13] is a resource used by the mammographic image analysis research community. This public database is a joint effort of American institutions (Massachusetts General Hospital, Wake Forest University and Washington University School of Medicine in St. Louis) and contains 2,620 cases, freely available on the Web.

2. *Pre-processing of Image*

The pre-processing stage involves techniques like K-means algorithm, Canny's filter, Hough's transform, erosion operator and histogram equalization which are necessary to find the orientation of the mammogram, to remove the noise from the image and enhance the quality. Digital mammograms are medical images analyzed by radiologists that are difficult to be interpreted thus pre-processing stage is used to remove the unrelated and surplus parts in the background of the mammogram. Hence, this phase is needed in order to improve the image quality and to make the segmentation results more accurate.

3. *Segmentation*

The segmentation stage helps in identifying Region of Interest (ROI), i.e. the breast regions where there is a larger possibility of being masses. Hence segmentation is performed using Cellular Neural Networks (CNN). According to chua et al. [10], standard CNN architecture consists of an M × N rectangular array of cells C (i, j) with Cartesian coordinates (i, j), i = 1, 2,…M, j = 1, 2…N. Thus from the mammogram the region of interest containing masses is been segmented which is helpful for the feature extraction stage to detect the masses in a specific region of interest.

4. *Feature Extraction*

Feature extraction is the fourth step in the proposed methodology. To generate feature vectors to be used in the classification stage it is done by extracting the shape and texture features. The shape features are eccentricity, circularity, compactness, circular disproportion and circular density and texture features are Ripley's K function, Moran's and Geary's indices.

- Ripley's K function:

Ripley's K function, in its local form, is computed by choosing a center i, where the occurrence of pixels j with equal gray level, for different radius(r) values, is examined. A is the area of the sample, n is the total number of points in the sample and d is a function that returns 1 if the distance dij between the points i and j is smaller than the radius r , or 0 otherwise:

$$K_i(r) = \frac{A}{n}\sum_{i \neq j} \delta\,(d_{ijr})$$

Here, each gray level is analyzed separately from the others to determine whether an event occurs or not within the specified distance r. Thus, the number of elements in the feature vector obtained through K(r) is given by the number of gray levels present in the image multiplied by the desired number of radii.

- Moran's and Geary's Indices:

Moran's index (I) is described by:

$$I = \frac{n}{W}\left(\frac{\sum i \sum j W_{ij} Z_i Z_j}{\sum i Z_i^2}\right) \qquad \text{for } i \neq j$$

where n is the number of observations, wij is the element in the proximity matrix for the pair i and j, W is the sum if the weights of the proximity matrix, zi and zj are the deviations in relation to the average, that is, zi = xi-x´ and zj= xj-x´.

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 8, August 2013

Geary's index (G) is given by:

$$G = \frac{(n-1)\sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}(X_i - X_j)^2}{2(\sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij})\sum_{i=1}^{n} Z_i^2} \quad \text{for } i \neq j$$

where xi and xj are the values of the variable of interest in the areas i and j, x is the average of the variables of interest in all areas of the sample, n is the number of areas in the sample, and wij is the element belonging to the proximity w and zi=xi-x´. At the end of this stage shape and texture features are extracted with the help of geostatistical functions and this makes detection of masses restricted to suspected regions of the breast.

5. *Classification*

In this final stage, classification of the candidate regions is done into mass or non-mass. For classification, two models are used; Support Vector Machine (SVM) and Probabilistic Neural Network (PNN) in this project work. Later, comparison is carried out between SVM and PNN in terms of specificity and sensitivity through ROC (Receiver Operating Characteristic) curve. The PNN classifier construction procedure is summarized as follows.

Step 1: The most representative neuron for each class from all the training samples is selected.

Step 2: Using all the selected representative neurons, construct a probabilistic neural-network classifier. Classify the training samples in each class and then compute the classification error rate, which is defined as the ratio of the number of misclassifications to the number of training samples in each class.

Step 3: Select one additional representative neuron using the neuron importance evaluating and ranking procedure Step 2 for classes that the requirement on classification error rate is not satisfied.

Step 4: Goto Step 2 until the requirement on classification error rate of all classes is satisfied. If the training samples are poor, the required classification accuracy might not be met even if all the training samples are used to construct the pattern layer. If this is the case, a higher classification error rate will be used.

Because only the most important neuron is selected at every step, the above procedure is capable of selecting a fairly small PNN with satisfactory classification accuracy.
The shape and texture features are been accessed by the PNN classifier to train some sample images. Then the trained images aids in testing the mammograms for classifying them into masses and non-masses. Therefore, according to results Probabilistic Neural Networks (PNN) reached a high accuracy as compared to SVM.

V. RESULTS

The snapshots of the implemented system are shown in the figures 4-9 below.

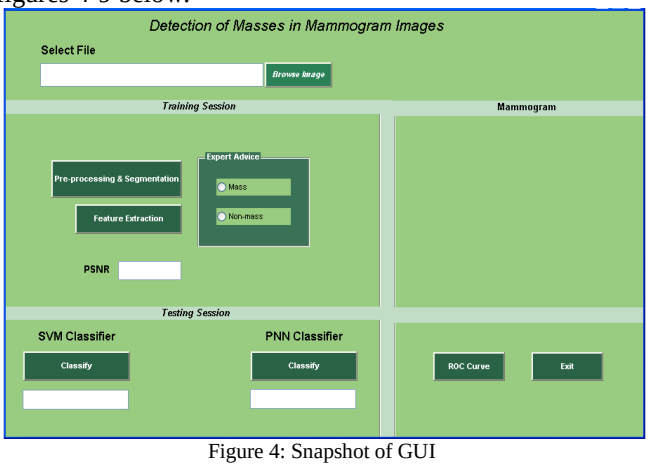

Figure 4: Snapshot of GUI

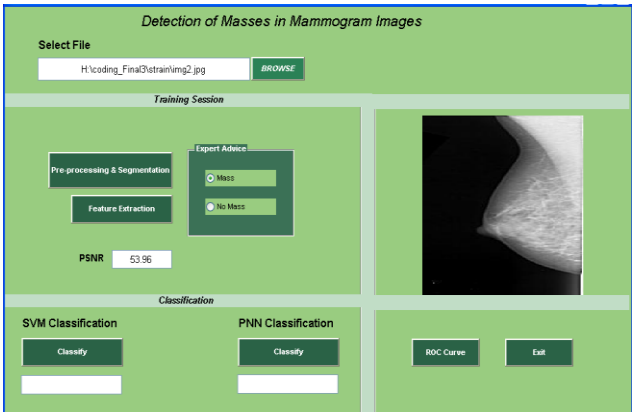

Figure 5: Browsing an Mammogram Image

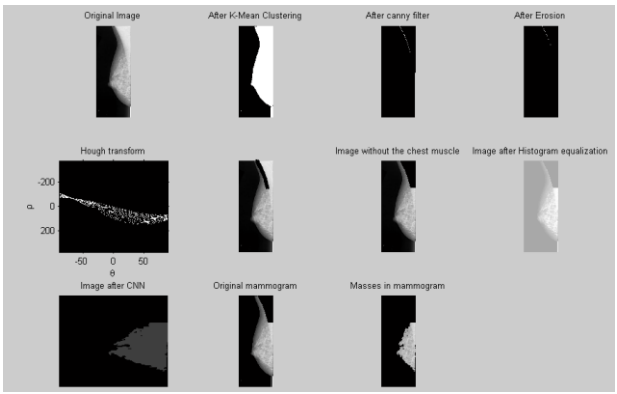

Figure 6: Pre-processing and Segmentation steps

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 8, August 2013

Sl.No	Features	Input Values to the Classifier
1	Eccentricity	0.383053
2	Circularity	2.24941
3	Compactness	0.000119718
4	Circular disproportion	0.0109416
5	Circular density	25503.1
6	Ripley's K Function	0.0156843
7	Moran's Index	47.3842
8	Geary's Index	0.219512

Figure 7: Shape and Texture features extracted

Features	Sensitivity (%)	Specificity (%)	Accuracy (%)
Shape	78.55	61.45	63.95
Ripley	87.25	56.01	60.58
Moran	75.65	55.32	58.29
Geary	69.36	63.0	63.93
Moran+Geary	78.57	70.17	71.43
Ripley+ Moran+Geary	**87.5**	**75.0**	**81.25**

Table 2: Results based on Shape and Texture Features

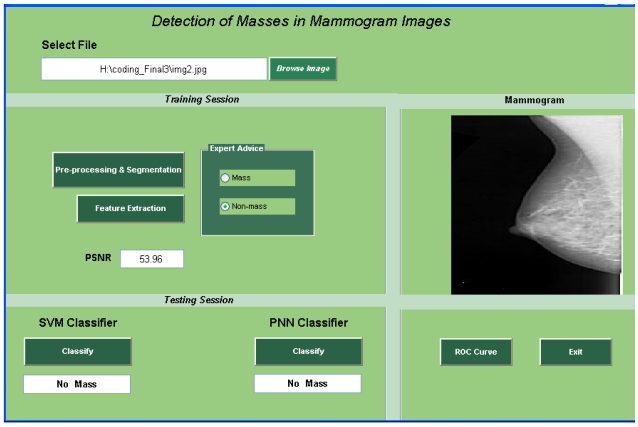

Figure 8: Classification based on SVM or PNN classifier

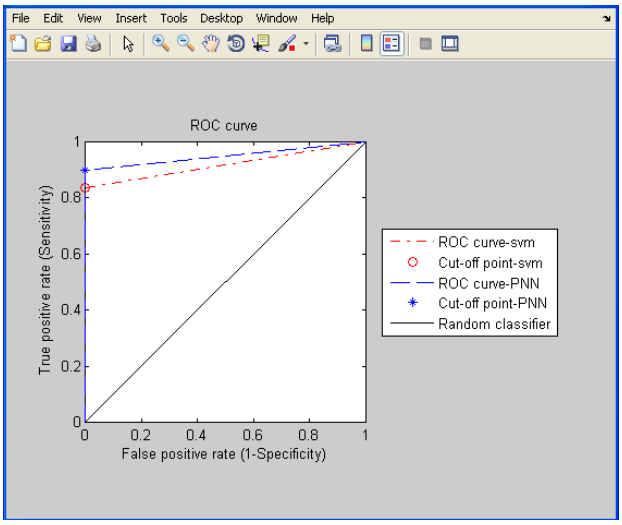

Figure 9: Comparison between SVM and PNN classifier based on ROC curve

VI. CONCLUSION

The mammographic detection of masses using CNN, Geostatistic functions and SVM & PNN classifiers has achieved a good performance. The results indicate that the combination of shape descriptors using geostatistic functions such as Ripley's K function, and Moran's and Geary's indices provides a good tool to characterize regions suspect of containing masses. The number of features required for higher efficiency using Cellular Neural Networks (CNN) and geostatistic functions has been optimized. The combination of geostatistic functions results in higher accuracy. When the performance was compared using ROC curve between SVM and PNN, relatively PNN achieved a high accuracy than SVM classifier.

REFERENCES

[1] N.C.I. (NCI), Cancer Stat Fact Sheets:Cancer of the Breast,available at http://seer.cancer.gov/statfacts/html /breast.htmlS, 2010.
[2]http://www.rxlist.com/collection-ofimages/breast_ anatomy_ picture/ pictures.
[3] Preet K. Dhillon, Breast Cancer Factsheet, 2003.11.11
[4] R.O. Duda, P.E. Hart, Pattern Classification and Scene Analysis, Wiley-Inter- science Publication, New York, 1973.
[5] I. El-Naqa, Y. Yang, M.N. Wernick, N.P. Galatsanos, R.M. Nishikawa, A support vector machine approach for detection of microcalcifications, IEEE Trans. Med. Imaging 21 (2002) 1552–1563.
[6] Wener Borges Sampaio, Edgar Moraes Diniz , Aristofanes Correa Silva, Anselmo Cardoso de Paiva , Marcelo Gattass, Detection of masses in mammogram images using CNN, geostatistic functions and SVM, Computers in Biology and Medicine 41 (2011) 653–664
[7] S. Haykin, Neural Networks: A Comprehensive Foundation, Prentice-Hall, Inc., 1999.
[8]Donald F. Specht, Probabilistic Neural Networks: LockheedMissiles & Space Company, Inc., Neural Networks, Vol. 3. pp. 109 118, 1990.
[9] M. Heath, K. Bowyer, D. Kopans, Current Status of the Digital Database for Screening Mammography, Digital Mammography, Kluwer Academic Publishers, 1998, pp. 457-460.
[10]L.O. Chua, T. Roska, Cellular Neural Networks and Visual Computing: Foundations and Applications, Cambridge University Press, New York, NY, USA, 2002 ISBN 0-521-65247-2.
[11]K. Z. Mao, K.C. Tan, and W. Ser, Probabilistic Neural-Network Structure Determination for Pattern Classification, IEEE Transactions On Neural Networks, Vol. 11, No. 4, July 2000 1009
[12]S. Agarwal, T. Graepel, S. Har-Peled, R. Herbrich, D. Roth, Generalization bounds for the area under the ROC curve, J. Mach. Learn. Res. 6 (2005) 393–425.
[13]http://marathon.csee.usf.edu/Mammography/DDSM/thumnails/normals/ normal_02/overview.html