A Research on Detection and Classification of Breast Cancer using k- means GMM & CNN Algorithms

S. Shamy, J. Dheeba

Abstract--- Breast cancer, is a type of cancer that affects women in larger number in the world. Medical advances on all fronts to improve the care of patients and defeat this disease of the century. Because of this, it is essential that several disciplines continue to make their contribution and particularly data mining or artificial Intelligence. The classification of breast cancer is a medical application that poses a great challenge for researchers and scientists. Recently, the neural network has become a popular tool in the classification of cancer datasets. The proposed method consists of three steps: The first step is to find region of interest (ROI). The second step is texture feature extraction of ROI and optimization of features using optimized feature selection algorithm.. The third step is classification of detected abnormality as benign or malignant using Convolutional Neural Networks (CNN). The proposed method was evaluated using Mammographic Image Analysis Society MIAS) dataset. The proposed method has achieved 95.8% accuracy.

Keywords--- Breast Cancer Classification, Convolutional Neural Network (CNN), K-means based GMM Algorithm, Medical Image Processing, Mammographic Images.

1. INTRODUCTION

Breast cancer is sort of malignant growth beginning from breast tissue, most commonly from the inner lining of milk ducts or the lobules of breast and then metastasizes to other areas of the body. In India, more than 100,000 women are recently determined with breast cancer every year; and have overtaken cervical cancer to turn into the main source for death among ladies in metropolitan urban areas. Breast Cancer is second most common cancer all over world [2]. Over 60% of breast cancers are detected in the complex stage and hence mortality from breast cancer are also high [3]. Hence premature detection of breast cancer is necessary in successful treatment and in dipping the number of deaths caused by breast cancer. Research related to the finding of breast cancer has improved for the duration of the final decade. Alto et al. [4] have chosen extracting descriptors of texture, shape and sharpness of the edge. Those related to pixel intensity, shape and texture were fused by Tao et al. [5], in order to find the tumors related to that contained in the ROI query and classify it as benign or malignant. For enhanced image resemblance, Zheng et al. [6] proposed a system that provides further interaction with the user; system in which, the latter is asked to assess the nature of the spiked tumor of the query image so that the system only looks for matches with similar degrees of speculation. This work was subsequently enhanced by removing from the search base the ROIs that gave the worst similarity scores [7].

Narvaez et al. [8] have explained a method that begins by merging the shape and texture descriptors extracted on the two incidences of the breast to discover the best matches, which images are then used to annotate the query ROI. Liu et al. [9] have on their side developed an image search based on a hash function to produce a diagnosis for the tumors contained in the ROIs queries. More specifically, a hash function inspired by graph theory and named anchor [10] was used to compress two descriptors, namely the SIFT histogram and the GIST into binary codes; finally the search for similarity was made in the Hamming space, Hang and Asoke [11] proposed a Multilayer Perceptron (PMC) as a classifier for diagnosing breast cancer. As a first step, a variable selection step is performed on the data using Genetic Programming (GP). This demand has led to the development of several techniques to detect tumor in early stages. Over the years, various researches have been done for the classification/detection of breast cancer. In this study, an efficient and automatic mammogram image classification system based on multiple modalities of features is proposed. In order to classify the tumor image accurately, the color and texture features must be extracted effectively. To provide assistance to the medical, robust and reliable diagnosis, neural networks can be a powerful tool for distributed diagnosis. In this paper, we tested the performance of the neural networks based on the MIAS Database. This paper is structured as pursues: Section 1 focuses the introduction & literature survey. Section 2 explains the methodologies, Section 3 deals with results and discussion and Section 4 shows the conclusion of the work.

2. PROPOSED WORK

The proposed system involves the following stages

Image Acquisition: The Mammography Image Analysis Society (MIAS) is an organization of UK research groups which have fashioned a digital mammography database. The images are in gray scale file format (PGM – Portable Gray Map).

The original MIAS Database (digitized at 50 micron pixel edge) has been reduced to 200 micron pixel edge and clipped/padded so that every image is 1024 pixels x 1024 pixels known as the mini-MIAS database. We have used mini-MIAS database as it contains complete information about abnormalities of each mammographic image.

Revised Version Manuscript Received on August 14, 2019.

S. Shamy, Research Scholar, Department of Computer Application, Noorul Islam Center for Higher Education, India. (e-mail: shamyshabeer@gmail.com)

J. Dheeba, Associate Professor, School of Computer Science and Engineering, Vellore Institute of Technology, India. (e-mail: dheeba.jacob@gmail.com)

Image preprocessing: In preprocessing the noises such as salt and pepper noise, label of mammogram image and black background in a mammogram are removed [7].

ROI selection: During the region of interest (ROI) selection the tumor region is identified based on highest intensity point and this region will be allocated for feature extraction.

Feature extraction and selection: Features such as energy values of wavelets and mean and standard deviation are calculated using gray level co-occurrence Matrix (GLCM) based texture analysis. The selection is performed by wrapper based GA algorithm.

Classification: Based on the selected features, the suspicious regions are classified using CNN algorithm. These main stages are illustrated in Fig.1

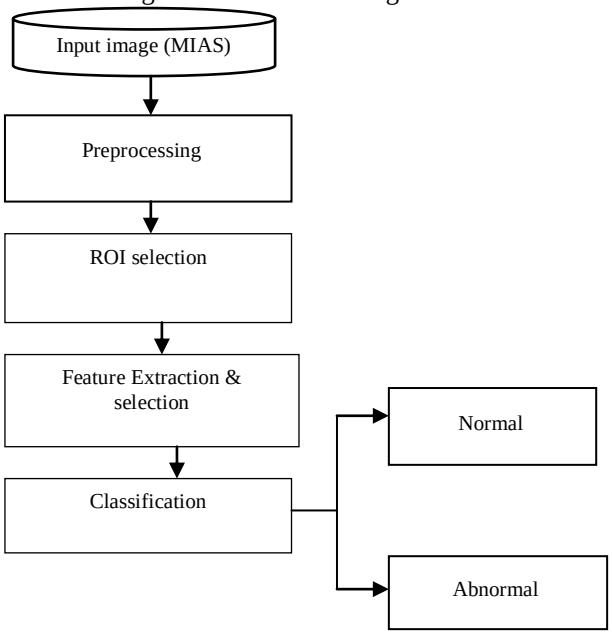


Figure 1: overall stages of proposed work

Pre-Processing

In preprocessing, the mammogram image is processed to remove noise and other abnormalities present. The mammogram is preprocessed using Adaptive median filter and bilateral filter. The Adaptive median filter is a nonlinear digital filtering technique that is used to remove noises such as impulse noise. The median filter does not perform well if the noise is > 0.2 dB but it clearly reduces the noise. In bilateral filter image intensity is highly improved, no more complex they preserve edges bilateral filters can be applied to color images.

ROI Selection

In this stage, the output from pre-processing stage is used as input to find the ROI which is done using K-means based GMM segmentation technique. K-means is iterative unsupervised clustering algorithm. Each cluster is characterized by its center point K-means finds a local minimum of the cost function and converges. Euclidean distance metric is used as Dissimilarity measure to find distance between pixel and centroid of each class. This is soft clustering algorithm. Each cluster is considered as a generative model with mean and variance. Mixture models are to estimate the parameters of probability distribution like mean and variance. The basic block diagram of K-means based GMM is shown in figure 2.

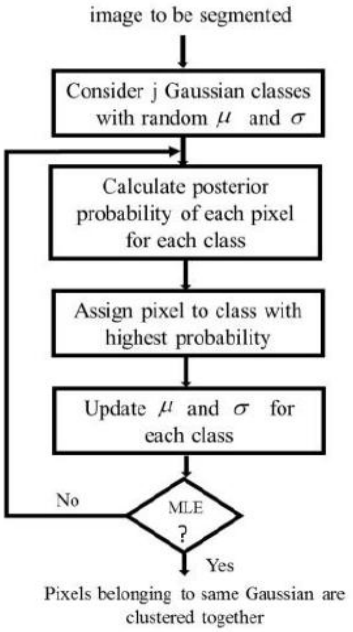


Figure 2: block diagram of K-means based GMM segmentation technique

Feature Extraction

In this stage features are extracted using texture analysis

Texture analysis of mammograms helps to identify texture feature information about the spatial distribution of tonal variations and describes the pattern of variation in gray level values in a neighborhood. Gray Level co-occurrence Matrix (GLCM) is used to extract texture information from images. The GLCM characterizes the spatial distribution of gray levels in an image. The features that are used for classification are: mean which is a measure of average intensity and standard deviation which is measure of average contrast submission. The features properties are described in table 1.

Table 1: Properties of Texture and co-occurrence matrix

Energy = $\sum_{i=0}^{N_p-1}\sum_{j=o}^{N_p-1} P^2(i,j)$	Contrast = $\sum_{i=0}^{N_p-1}\sum_{j=o}^{N_p-1}(i,j)^2\, p(i,j$
Correlation = $\sum_{i=0}^{N_p-1}\sum_{j=o}^{N_p-1}\frac{(i-\mu)(j-\mu)\,p(i,j}{\sigma_i\sigma_j}$	Homogeneity = $\sum_{i=0}^{N_p-1}\sum_{j=0}^{N_p-1}\frac{p(i,j)}{1+\lvert i-j\rvert}$
Mean: $\mu_i = \frac{1}{N}\sum_{j-1}^{N} f_{ij}$	SD: σ_i = $\left(\frac{1}{N}\sum_{N}^{1}\left(f_{ij}-\mu_i\right)^2\right)^{\frac{1}{2}}$
Skewness: γ_i = $\left(\frac{1}{N}\sum_{N}^{1}\left(f_{ij}-\mu_i\right)^3\right)^{\frac{1}{3}}$	Entropy: ε_i = $-\sum_{j-1}^{N}\left(f_{ij}-\log f_{ij}\right)$

Feature Selection based wrapper based method

Feature selection by GA and Optimization of feature selection is done by Wrapper Method/ Hill Climbing, The wrapper method is a random search technique used for selecting the relevant subset features from the available feature set. The Genetic algorithm is incorporated as a wrapper based model in the proposed algorithm. In each generation the population is evaluated and tested with the termination algorithm. If it fails the crossover, mutation and fitness computation steps are repeated

Genetic algorithm

In order to solve optimization problems dependent upon natural selection, GA uses Darwinian criteria of "Survival of the fittest", for the optimization of population evolution. In order to find out the best environment and to increase the efficacy of the set of probable solutions, natural selection is utilized. Genetic algorithm was presented by Holland in the early 1970's. Robustness in genetic algorithm, makes its structural functionality to be suitable for numerous diverse search problems .The most prominent dissimilarity between the genetic algorithms and numerous conventional optimization approaches is that the genetic algorithm utilizes a population of points at onetime, conflicting to the single point method by conventional optimization techniques.

For better performance of genetic algorithm, two conditions are to be satisfied i.e, it needs a genetic representation and fitness function of the solution domain. Once these functions are defined, GA randomly tries to initialize the population and it is improved by the application of genetic algorithm operators such as selection, crossover and mutation.

Algorithm for GA

Step1: Initial population of individuals is chosen first

Step2: Fitness function evaluation of the population

Step3: generation process is repeated until termination criteria is reached (time limit, sufficient fitness achieved etc.):

- The best-fit individual is selected for reproduction.
- New individuals are created through crossover and mutation operations.
- New individual's fitness evaluation
- Replacing the least-fit population with new individuals

The proposed methods for generating feature subset can reduce the curse of dimensionalities to a great extent.

Classification

In this stage the features obtained from previous stage are converted to feature vectors. These feature vectors are used for differentiating between a micro calcification and a circumscribed mass and they are also further classified into benign or malignant case as shown in fig.2. Classification is done using a CNN classifier.

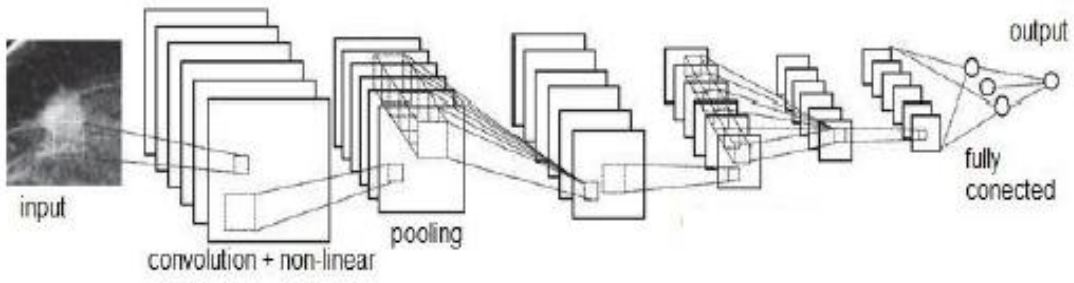


Figure 3: Convolutional Neural Networks

Convolutional Neural Network (CNN) is feedforward neural network developed by Kunihiko Fukushima in 1980 and improved by Yann LeCan et al. in 1998. A CNN is composed of 6 types of layers: an input layers, a convolutional layer, a non-linear layer, a pooling layer, fully connected layer, and an output layers. Fig. 3 illustrates a traditional CNN architecture. Convolutional Neural Networks (CNN) are one of the most remarkable approaches of deep learning, in which multiple layers of neurons are formed in a robust manner. They have shown that they are capable of demonstrating an impressive generalization capability on large data sets with millions of images. These results come mainly from the particular architecture of CNNs that takes into account the specific topology of tasks related to the field of computer vision that exploit two-dimensional images. Other dimensions can also be taken into account when it comes to color images with multiple channels.

3. EXPERIMENTAL RESULTS

Presented below are the results of the proposed method carried out on mini-MIAS dataset. Fig.4(a) shows the original image and Fig.4(b) and 4(c)shows the preprocessing step which involves label removal. Fig.5(a) shows the detected abnormality which will be the ROI for feature extraction. The texture features are extracted for the ROI, the result of feature extraction is shown in table 2. Fig 6(a) shows the first section of resnet and fig 6(b) shows the first convolutional layer weights. Fig 6(c) shows the confusion matrix of the proposed work. The accuracy of proposed method is 95.8%

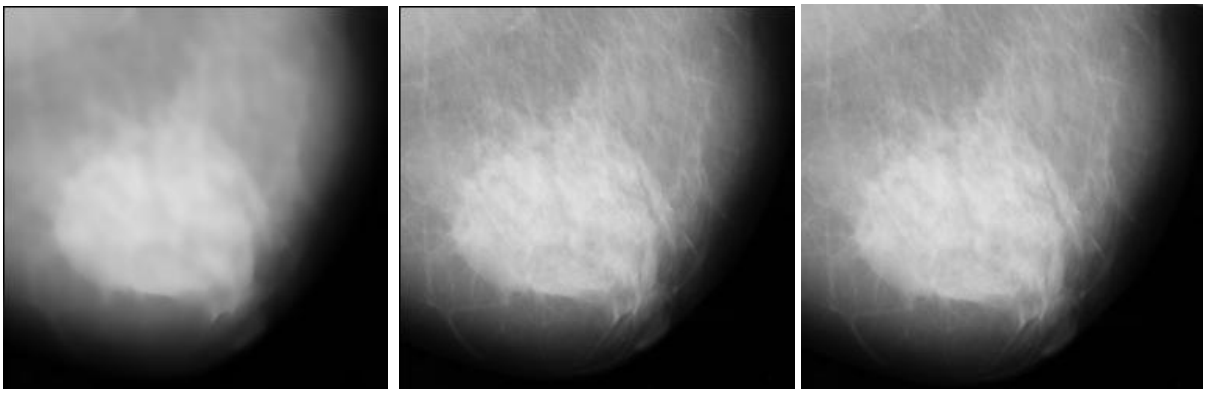
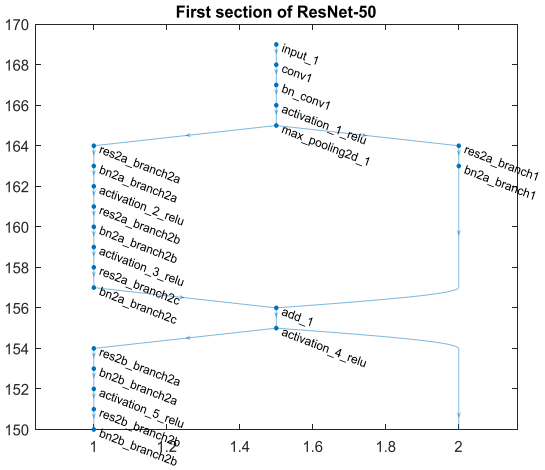


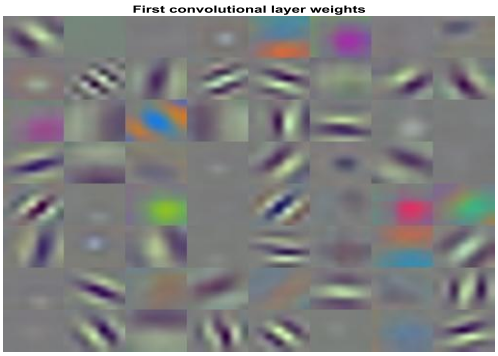
Figure: 4: (a) Input image, Comparison of both (b) adaptive median filter image (c) bilateral filter image

Figure 5: segmented images of k- means based GMM algorithm

Table 2: Texture and co-occurrence matrix results

Features	Texture and co-occurrence matrix							
Images	Co	Cor	E	Ho	M	Sd	S	En
1	0.25491	0.988181	0.290193	0.922654	0.003906	0.011445	10.28146	1.696392
2	0.210345	0.985636	0.19678	0.941828	0.003906	0.016421	10.14462	1.390808
3	0.205686	0.984743	0.197149	0.946424	0.003922	0.015425	10.262	1.447847
4	0.220095	0.98512	0.200374	0.943258	0.003906	0.016679	10.01903	1.394749
5	0.21385	0.984853	0.187032	0.943495	0.003922	0.01597	10.28368	1.435335





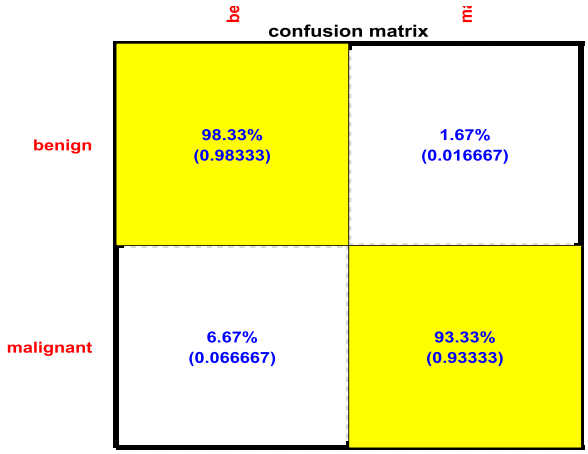


Fig. 6: Over all outputs of CNN algorithm (a) first section of resnet (b) the first convolution layer weights (c) confusion matrix

4. CONCLUSION

The proposed algorithm is tested on a real life problem, the Breast Cancer Diagnosis problem. The objective of this study is to create an effective tool for building neural models to help us making a proper classification of various classes of breast cancer. Neural network approach drove by the learning algorithm works well, in terms of accuracy. Using this model, an automated classification of various types of breast cancer was performed by avoiding the question of the expert concerning the recognition of cancer required, improving the identification of breast cancer classification. Through the results analysis it was found that the proposed model reduces significantly the computing time and the solutions quality is improved significantly.

REFERENCES

1. Survey by Indian cancer society, Indian Cancer Society (2018)
2. Padmanabhan, S., Sundarajan, R.: Enhanced Accuracy of Breast Cancer Detection in Digital Mammograms using wavelet analysis. IEEE Trans. Imag. Proc. (2012).
3. Spandana, P., Rao, K.M.M., Jwalasrikala,

4. J.: Novel Image Processing Techniques for Early Detection of Breast Cancer. In: Matlab and Lab View Implementation. IEEE Pointof- Care Healthcare Technologies (PHT), Bangalore, India, pp. 16–18 (2013)
5. H. Alto, R. M. Rangayyan and J. L. Desautels, Content-based retrieval and analysis of mammographic masses. Journal of Electronic Imaging, Vol 14, No.2,:023016– 023016, 2005
6. Y. Tao, S. Lo, M.T. Freedman and J. Xuan, A preliminary study of content-based mammographic masses retrieval. In Medical Imaging, 65141Z. International Society for Optics and Photonics, 2007.
7. B. Zheng, A. Lu, L. A. Hardesty, J. H. Sumkin, C. M. Hakim, M. A. Ganott and D. A. Gur, A method to improve visual similarity of breast masses for an interactive computer-aided diagnosis environment. Medical Physics, Vol 33, No.1, pp: 111–117, 2006.
8. C. Wei, Y. Li and P. Huang, Mammogram retrieval through machine learning within bi-rads standards. Journal of Biomedical Informatics, Vol 44, No.4, pp: 607–614, 2011.
9. Narvaez, G. Diaz and E. Romero, Multi-view information fusion for automatic bi-rads description of mammographic masses. In SPIE Medical Imaging, 79630A. International Society for Optics and Photonics, 2011
10. J. Liu, S. Zhang, W. Liu, X. Zhang and D. Metaxas, Scalable mammogram retrieval using anchor graph hashing. In 2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI), pp; 898–901, 2014.
11. W. Liu, J. Wang, S. Kumar and S. Chang, Hashing with graphs. In Proceedings of the 28th international conference on machine learning (ICML-11), pp; 1–8, 2011
12. G. Hong and N. Asoke, Breast cancer diagnosis using genetic programming generated feature. Pattern Recognition, 2006, vol. 39, no 5, p. 980-987.
13. V. Vishrutha and M. Ravishankar, Early Detection and Classification of Breast Cancer, Springer International Publishing Switzerland 2015
14. Htet Thazin Tike Thein and Khin Mo Mo Tun, An approach for breast cancer diagnosis classification using neural network, Advanced Computing: An International Journal (ACIJ), Vol.6, No.1, January 2015
15. Majid Nawaz, Adel A. Sewissy, Taysir Hassan A. Soliman, Multi-Class Breast Cancer Classification using Deep Learning Convolutional Neural Network, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 9, No. 6, 2018
16. U. Baid, S. Talbar and S. Talbar, Comparative Study of K-means, Gaussian Mixture Model, Fuzzy C-means algorithms for Brain Tumor Segmentation, ICCASP/ICMMD-2016. Advances in Intelligent Systems Research. Vol. 137, Pp. 592-597
17. Josephine, Brain Tumor MRI Image Detection and Segmentation Using Genetic Algorithm, International Journal of Computer Sciences and Engineering Volume-6, Special Issue-2, March 2018
18. Muhamamdu Sathik Raja and K.VishnuLakshmi Genetic Algorithm Based Brain Tumor Detection and Segmentation, International Journal of Innovative Research in Advanced Engineering (IJIRAE), Issue 03, Volume 4 (March 2017).