International Journal of Applied Engineering Research
ISSN 0973-4562 Volume 10, Number 7 (2015) pp. 17643-17656

http://www.ripublication.com

Detection of Masses In Digital Mammograms Using K-Means and Neural Network

S. Julian Savari Antony[1], S.Ravi[2]
[1]*Research Scholar, Department of Electronics and Communication Engineering*
Shri J.J.T. University, Jhunjhunu, Rajasthan, India
e-mail: savari.sm@gmail.com
[2]*Faculty of Electrical Engineering, Selvam College of Technology*
Namakkal, Tamil Nadu, India

Abstract

Breast cancer is a serious public health problem in several countries. It presents an efficient computer aided mass classification method in digitized mammograms using Neural Network and K Means, which performs benign-malignant classification on region of interest (ROI) that contains mass. This paper presents a research on mammography images using K-means for detecting cancer tumor mass and micro calcification to help of initiated centroid values which have separated for different area of values. One of the major mammographic characteristics for mass classification is texture. Neural Network exploits this important factor to classify the mass into benign or malignant. The statistical textural features used in characterizing the masses are entropy, standard deviation, mean, variance and co-variance. It has used different data groups with more number of samples. In order to estimate the future method. The main aim of the method is to increase the effectiveness and efficiency of the classification process in an objective manner to reduce the numbers of false-positive of malignancies. The proposed technique shows better results in less time complexity (in Seconds).

Keywords: Mammographic Images, Neural network (NN), K means, Region of Interest, Texture Analysis

Introduction

Breast cancer is one of the most overwhelming reasons of the demise among women in the world and mammography image is quiet the most commonly used method for detecting breast cancer at early stage.However, radiologists can miss a significant portion of abnormalities. Some studies indicate that Computer Aided Detection systems (CADe) can deliver a second opinion to the radiologists and potentially

decrease the missed detection rate [1].A CADe system used in breast cancer screening programs is collected by two main steps: the identification of suspicious regions and the false positives reduction [2] . Algorithms for the False Positive Reduction (FPR) of suspicious signs of disease, can work either with one view or with multiple views [3]. Typically, the one view FPR is a two class's classification task in which each Region of Interest (ROI) can be classified as a mass or as normal breast tissue. A set of geometric and/or textural features have to be extracted and selected to train the classifier. Alternatively, template matching approaches can be used, comparing each extracted ROI with all the ROIs of a certain database using similarity measures or features vectors.

Neural networks have supported for breast cancer detection by several researchers. Various efforts to refine classification presentation have made, by a number of plans involving some means of choice between alternatives. Ensembles have been proposed as a mechanism for improving the classification accuracy of existing classifiers [4] providing that elements are diverse.

It designed a CAD system for breast mass detection on digital breast tomosynthes is (DBT) mammograms. Each mass candidate was segmented from the structured background, and its image features were extracted. A feature classifier was designed to differentiate true masses from normal tissues. The CAD system achieved a sensitivity of 85%, with 2.2 false-positive objects per case [5]

It proposed and evaluated the performance of a CAD algorithm in marking preoperative masses. First, Digitized mammograms were processed with an adaptive enhancement filter shadowed by a local border refinement stage. Test results showed that malignant masses were detected with the computer in 87% (135 of 156), 83% (130 of 156), and 77% (120 of 156) of the malignant cases at FPI rates of 1.5, 1.0, and 0.5 marks per mammogram, respectively [6]

They used a recursive median filtering technique that could be applied to images at a number of scales and orientations, giving a scale space description at pixel level. This technique was applied to mammography, in the detection of mass-like structures associated with speculated lesions. A sensitivity of 80% was achieved with 0.25 false positives per image [7]

Classification was another most important process in CAD system design. In [8] used improved local binary pattern operator for mass classification. The papers [9], [10], 11] used support vector machine with combination of different techniques for the classification of masses. Naïve bayes classifier [12], K means classifier, fuzzy C means clustering [13][14] are some of the common methods were used by the previous work. In [15] designed least square support vector machine which provided effective classification compared to other methods. An ANN was configured for a specific application, such as pattern recognition or data classification, through a learning process. Learning in biological systems involves adjustments to the synaptic connections that exist between the neurons [16].

They were used several methods and got from output in different kinds of accuracy, false positives and time complexity.

Proposed Method

The proposed system is following the process as k- means and neural network.

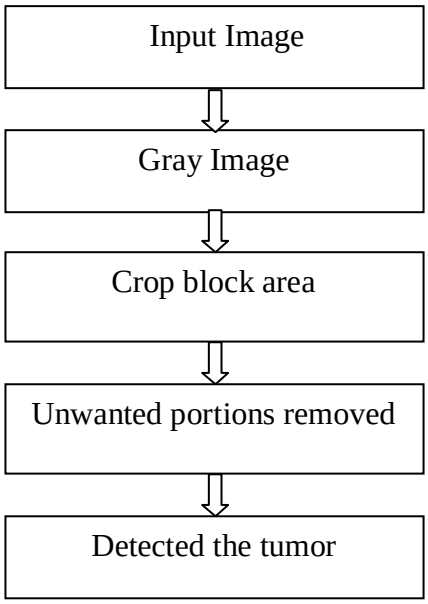


Figure 1: Flow chart showing the Main Steps in K- Means Algorithm

Image Pre-processing

Image pre-processing is a technique to analyze the image before start executing our process. In this technique by using the crop method the size of the original image is re-sized in the selected or exact area. For example the unwanted area of the real image is eliminated over here in shown fig.2

Image Enhancement

Next step to the image preprocessing is image enhancement. Image enhancement is the process by which we can able to improve the quality of digitally stored images. In enhancement method was improving the quality and value of the original image which is most important for further process in shown fig.2. which is mentioned by UPR.

K-Means Algorithm

Images were classified as cancerous or non-cancerous by best fit into a cluster and are assigned to that cluster. The K-Means algorithms were used for the purpose. The algorithm uses random seeds, i.e., points with random mean values to form lines to separate the classes. Next, the points within the delineated areas are analyzed, and their mean values are calculated. The means form the new seeds from which a new series of lines can be formed to separate the classes. This process is done repeatedly. The advantage of this method is that it has the potential to model complex target

functions with a small set of features. The clustering works based on the following equations:

$$D(i,k) = \| (Xk - Vi) \| 2 \text{ for } I \leq C, K \leq N, \quad (1)$$

$$Vi(l) = (\sum Ni\ Xi\ )/Ni \quad (2)$$

$$\prod_{k=1}^{n} max \left| V(l) - V(l\text{-}1) \right| \neq 0 \quad (3)$$

D(i,k) calculates the distances between each class, c is the number of clusters, N is the number of objects in the cluster and v determines the cluster center. The higher value of k results in smooth grouping. The method follows the usual steps to satisfy the primary objective: clustering all the image objects into K distinct groups. First, K centroids are defined, one for each group, being their initial position very important to the result. After that, it is determined a property region for each centroid, which groups a set of similar objects. The interactive stage of the algorithm is started, in which the centroid of each group is recalculated in order to minimize the objective function. This function, for K-means, is the minimum square method, calculated by tumor area. It has divided to centroid four values. The input image applied to k-means that have generated by initiated centroid value each images. To find the tumor area is using equation .3.

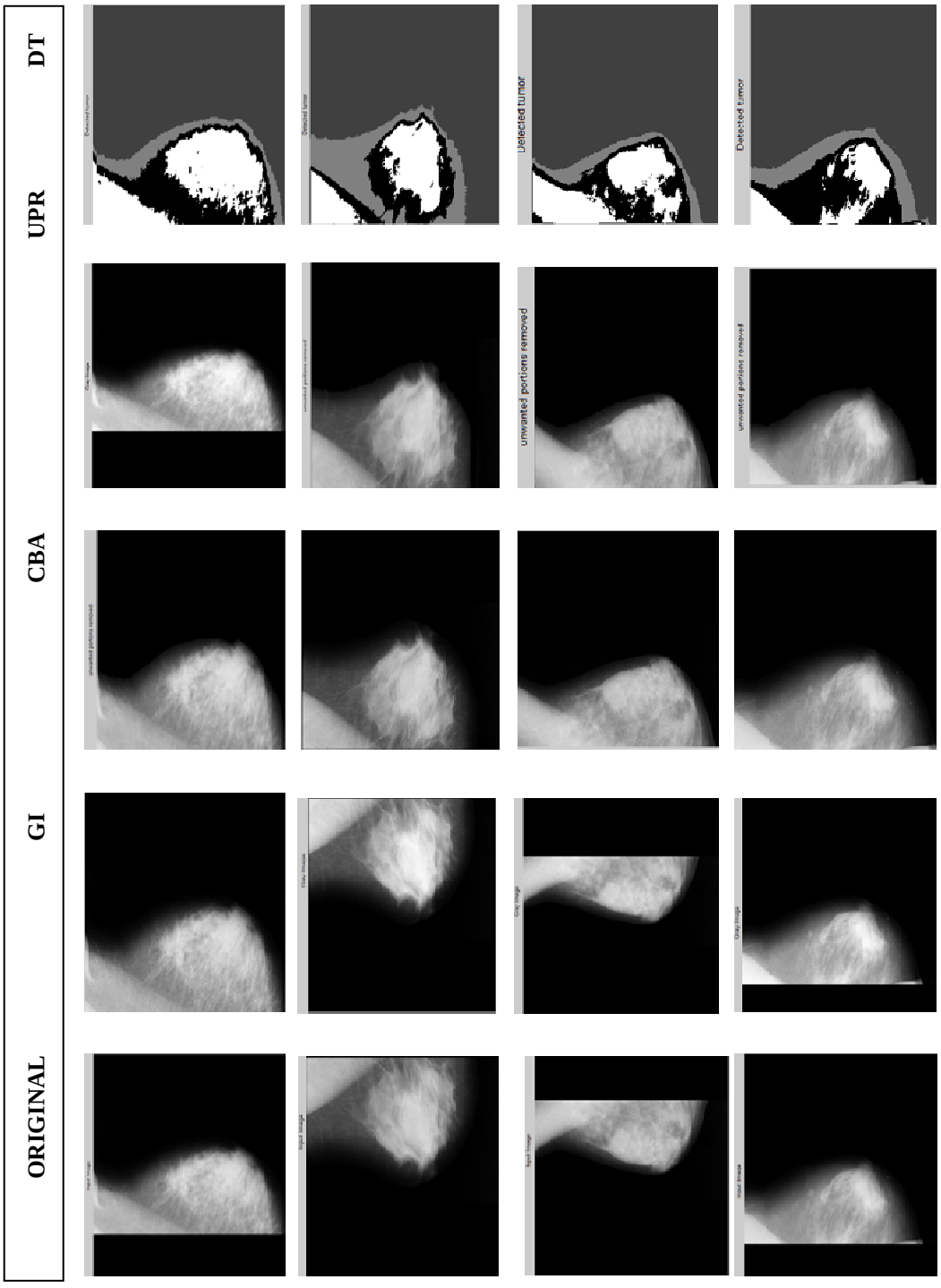
The K- Means Algorithm can be done by the following work flow. As a base step an image is taken from a particular portion as a gray image. Crop the block portion of gray image and remove the unwanted portion then we can easily able to detect the tumor portion exactly.

Result and Discussion

The perform mammogram classification. Each approach has been implemented using Matlab and evaluated for the efficiency of the classification. For the evaluation, It has used different data sets with more number of samples. In order to evaluate the proposed method, the Department of Defense (DoD) data base for breast cancer research program from Standford School of medicine, California and Mammographic Image Analysis Society (MIAS) data set and Digital Data set for Screening Mammography (DDSM) has been developed.

Table 1: Usage of database

Database	Number of samples	Number of testing images
DoD BCRP	750	12
MIAS	322	5
DDSM	2640	50



GI: Gray image, CBA: Crop Block Area, UPR: Unwanted portions Removed, DT: . Detected the Tumor

Figure 2: Detect the tumor using k-mean algorithm

The usage of data was taken from Table 1.The fig.2 shows that from the original image, initially gray image is obtained. Then, the required area has been cropped and shown as Crop block area. Also, the unwanted portions are removed which is shown above as UPR in order to obtain the actual cancer affected area. The final part of the figure shows that the tumor has been deducted as clearly visible.

Table 2: Detected tumor area using K-Means Algorithm

Images	ICV–I	ICV- II	ICV-III	ICV-IV
IMG001	44.4000	88.8000	133.2000	177.6000
IMG002	48.0000	96.0000	144.0000	192.0000
IMG003	44.8000	89.6000	134.4000	179.2000
IMG004	46.2000	92.4000	138.6000	184.4000
IMG005	43.8000	87.6000	131.40000	175.2000
IMG006	46.6000	92.8000	139.2000	185.6000
IMG007	47.0000	94.0000	141.0000	188.0000
IMG008	45.6000	91.2000	136.8000	182.4000
IMG009	46.2000	92.4000	138.6000	184.8000
IMG010	46.6000	93.2000	139.8000	186.4000
IMG011	45.4000	90.8000	136.2000	181.6000

ICV-Initiated centroid value

The Table 2: shows that the tumor has been deducted on 12 different images obtained from 12 samples by using K-means algorithm. It is clearly shown using initiated centroid value at four different levels such as I, II, III and IV.

Feature Extraction

Texture feature is used to quantity the surface variation of the image. This measurement helps to differentiate malignant or benign. Features. Segmentation output is given as input to feature extraction process. Spatial Gray Level Dependence is the joint probability of occurrence of gray levels i and j for the two pixels with a defined spatial relationship in an image.

Entropy

The entropy h can be described as a measure of the maximal amount of potential information given by the segmented ROI. Overall entropy of the image can be calculated as,

$$h = -\sum_{k=0}^{L-1} Pr_k \ (\log_2 Pr_k) \quad (4)$$

Where Prk is the probability of the k-th grey level, which can be calculated as Zk/M×N, Zk is the total number of pixels with the k-th grey level and L is the total number of the available grey levels in an ROI of size M×N.

Standard Deviation

The standard deviation σ is the estimate of the mean square deviation of grey pixel value p(i,j) from its mean value μ. Standard deviation describes the dispersion within the ROI. That is why the standard deviation feature is sometimes called the image dispersion. The standard deviation is defined as,

$$\sigma = \sqrt{\frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}(p(i,j)-\mu)^2} \quad (5)$$

The standard deviation is calculated directly using the instruction (sd), which is equivalent to applying (eq.5) on the segmented ROI.

Mean Value

The mean μ of the pixels values in the segmented ROI, estimates the value in the ROI in which central clustering occurs. In other words, it represents the average of all the pixels in the segmented ROI, as shown in the following equation,

$$\mu = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}p(i,j) \quad (6)$$

Where p(i,j) is the pixel value at point (i,j), in an ROI of size M×N. The instruction (mean2) can be used to compute the mean value of the foreground ROI only after neglecting the black background pixels from the calculation.

Variance

The variance [7] is a measure of how far a set of numbers is spread out. It is one of several descriptors of a probability distribution, describing how far the numbers lie from the mean (expected value). In particular, the variance is one of the moments of a distribution. In that context, it forms part of a systematic approach to distinguishing between probability distributions. While other such approaches have been developed, those based on moments are advantageous in terms of mathematical and computational simplicity. Mathematically variance is given by

$$\sigma^2 = \frac{1}{MN-1}\sum_{(i,j)\epsilon W}\left(p(i,j)-\frac{1}{NM-1}\sum_{(i,j)\epsilon W}p(i,j)\right)^2 \quad (7)$$

Variance filter can be utilized to determine edge position in image processing.

Covariance

In statistics, covariance [4, 5, 7] is a measure of how much two random variables change together. If the greater values of one variable mainly correspond with the greater values of the other variable, and the same holds for the smaller values, i.e. the variables tend to show similar behavior, the covariance is a positive number. In the opposite case, when the greater values of one variable mainly correspond to the smaller values of the other, i.e. the variables tend to show opposite behavior, the covariance is negative. The sign of the covariance therefore shows the tendency in the linear relationship between the variables. Mathematically covariance between two variable *x* and *y* is given as

$$Cov(N,M) = \frac{1}{T}\sum(N_i-\bar{N})\,(M_j-\bar{M}) \quad (8)$$

Covariance based filtering techniques can be used in edge sharpening, appearance based face detection and image orientation information etc

Data Values

The data values is summation of all values (entropy, standard deviation, mean value**,** variance, covariance and TA) divided by hundred.

$$D = \frac{(E+S.D+ME+V+C.V+TA)}{100} \quad (9)$$

Data values = (E+S.D+ME+V+C.V+TA)/100
= (0.4729+0.7139+1.4011+1.1911+0.0414+50516)/100
= 505.1982 (in refer to table3: IMG01)

Sample result of feature extraction process for 12 mammograms is shown in Table.3

Table 3: Sample Feature Values for Mammogram images

Images	E	SD	ME	V	CV	TA	Data	A
IM01	0.4729	0.7139	1.4011	1.1911	0.0414	50516	505.19	M
IM02	0.4720	0.6893	1.4922	1.3684	0.0467	55496	555.00	M
IM03	0.4700	0.7874	1.3453	1.2289	0.0135	50153	501.56	M
IM04	0.5731	0.6929	1.1665	0.8678	0.0177	51641	516.44	M
IM05	0.7553	0.7553	1.7232	2.3434	0.0397	32628	326.33	M
IM06	0.8943	0.7222	1.1659	1.5259	0.0449	36782	367.86	M
IM07	0.8979	0.7182	1.1021	1.3704	0.0249	37826	378.30	M
IM08	0.5675	0.8479	1.4915	1.6510	0.0277	44387	443.91	B
IM09	0.4906	0.7723	1.3480	1.1950	0.0185	50289	502.92	B
IM10	0.6500	0.8187	1.3839	1.4624	0.0376	44852	448.56	B
IM11	0.7694	0.7315	1.1848	1.3008	0.0320	43116	431.20	B
IM12	0.5944	0.8016	1.3252	1.2709	0.0435	47763	477.67	B

E: Entropy, S.D: Standard Deviation, ME: Mean, V: Variance, C.V: Co-Variance, T A:Tumor Area in Sq. mm, A:Assessment, M:Malignant, B: Benign, A: Assessment

The table3: shows that the Sample Feature Values for Mammogram images. It depicts the entropy, standard deviation, mean, variance, co-variance, tumor area in sq. mm. The final column shows that whether the classification is malignant or benign.

Neural Network

Any classification method uses a set of features or parameters to characterizeeach object, here these features should be relevant to the task at hand. There are twophases of constructing a classifier. First is the training phase, in which atraining set is used to determine how the features are to be weighted and combinedin order to classify the objects. Secondly, in the application phase, the weights obtained from the training set are applied to a set of new objects for classification. To obtain a better classification rating, a classifier based on neural networks was designed. The architecture of the network) is a multi-layered one where the nodes in a layer are fully connected to the

nodes in the next layer. The input layer contains the fractal feature values, such as fractal dimension and fractal signature. The hidden layer contains five nodes and the output layer has an output node. This neural network is trained using the back propagation algorithm.

Once the tumor is detected then it can be leveled by using neural network. Leveling is the process to determine the effected level of tumor. Once the pectoral image is extracted as the extraction image which is used to remove the pectoral muscle and to determine the tumor as in evil or in beginning stage. A set of twelve images are obtained and neural network operation is made. This shows the Sample Feature Values for Mammogram images. The fig.4 shows that initially, the pectoral image is obtained. Then extraction is carried out inorder to obtain the Extraction Image. Also, the pectoral removal takes place and final figure shows that the whether the tumor is malignant or benign in Fig.4. Detect the classification using neural network.

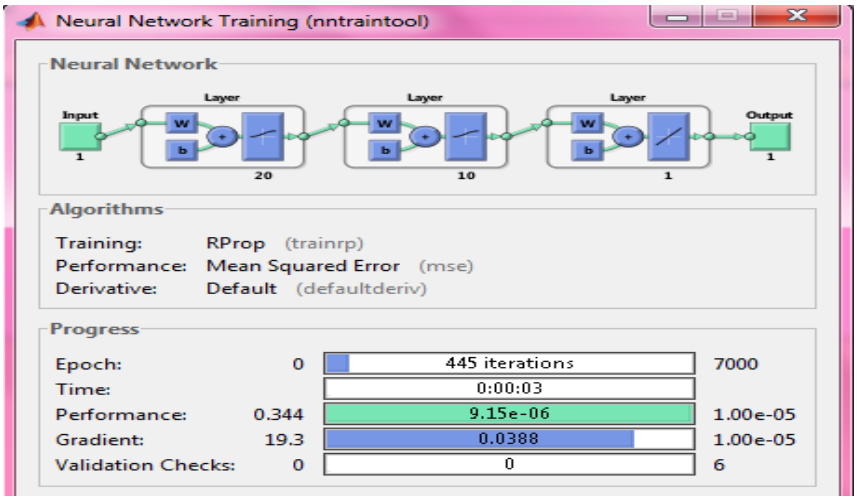
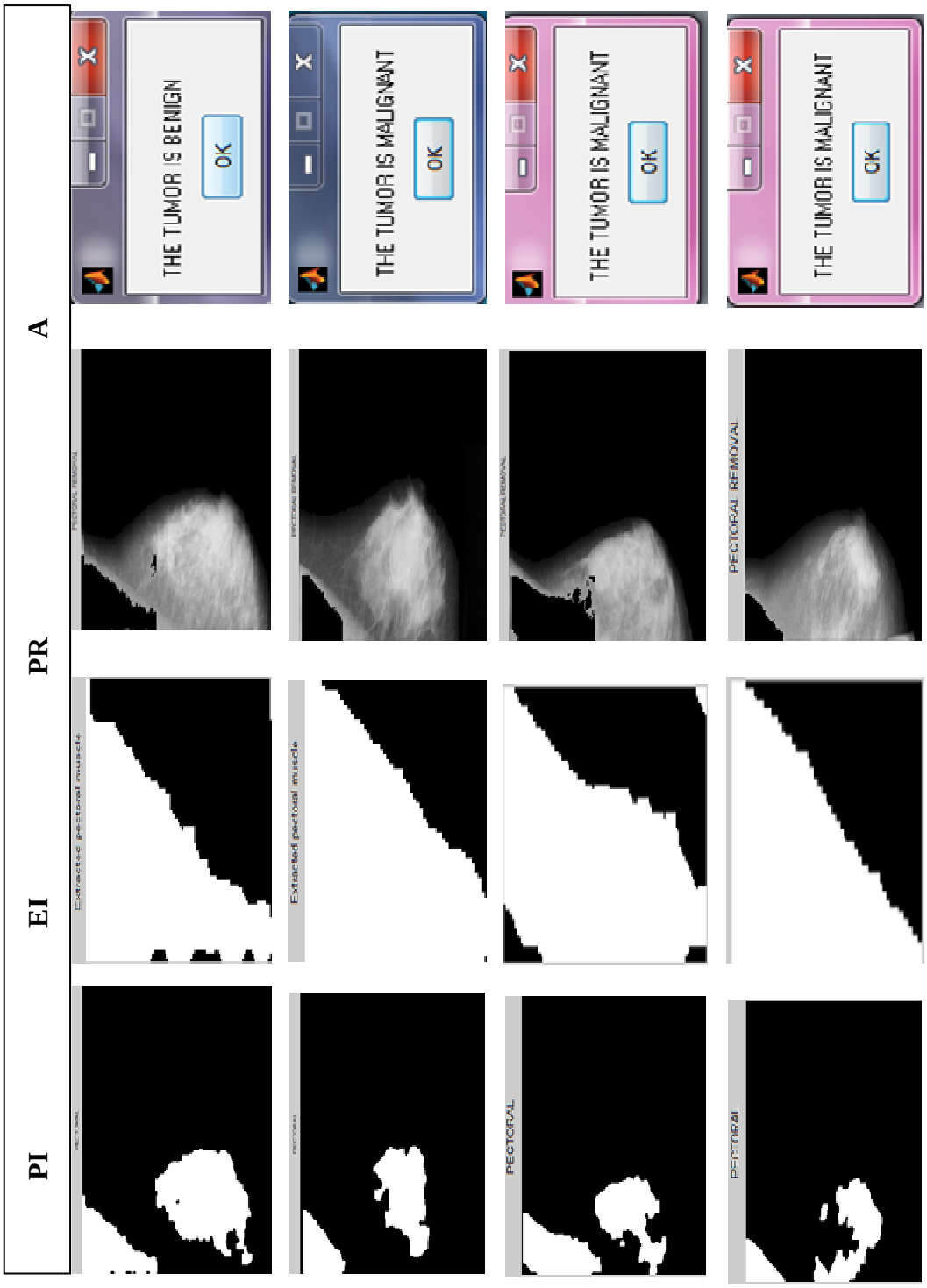


Figure 3: Neural Network Training

One of the most important and plus point of this system is operation time is of just 3 seconds only.



PI: Pectoral Image, EI: Extraction Image, PR: Pectoral Removal Muscle, A: Assessment.

Figure 4: Detect The Classification Using Neural Network

Table 4: Time Complexity

Image	Time in seconds	Over all Time Complexity
IMG001	0.392590	
IMG002	0.354468	
IMG003	0.386304	
IMG004	0.366614	
IMG005	0.368749	
IMG006	0.425379	0.373403
IMG007	0.379585	
IMG008	0.360682	
IMG009	0.344904	
IMG010	0.353756	
IMG011	0.399585	
IMG012	0.319585	

The table 4 shows time complexity which is obtained from 12 images. The second column shows time in seconds for 12 different images. The third column shows the overall time complexity which is obtained as 0.373403.

The accuracy of each and every three proposed techniques were shown in percentage and neatly represented in graphical manner as shown in fig 5. Accuracy is the ratio between the numbers of true positive results to that of the total number of samples have been taken. The accuracy of DoD method is 98.4 %, MIAS method is of 98.4472% and the DDSM method reaches 98.10606%.

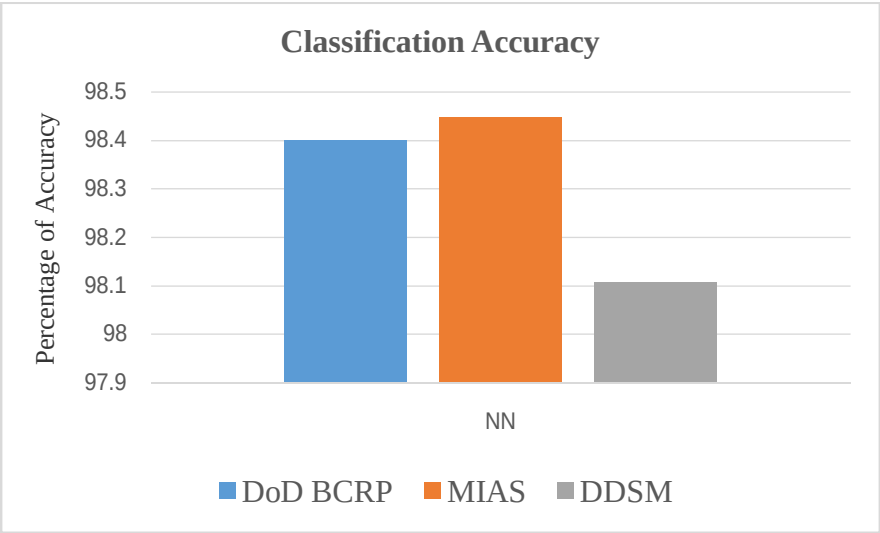


Figure 5: NN of classification accuracy

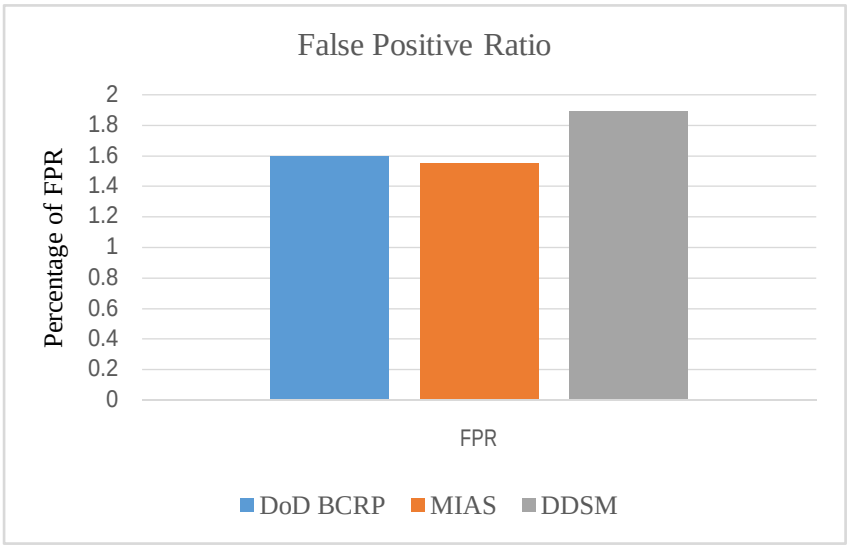


Figure 6: NN of false positives classification ratio

The false positive ratio is defined as the relation between numbers of false positive result to the total number of samples. In other words, it is the percentage difference from the accuracy which is having been shown in fig.6. DDSM method has the lowest false positives when compared to the other two methods. By calculating under this method, the false positive of DoD method is 1.6%, MIAS method is of 1.552795% and the DDSM method reaches 1.893939%.

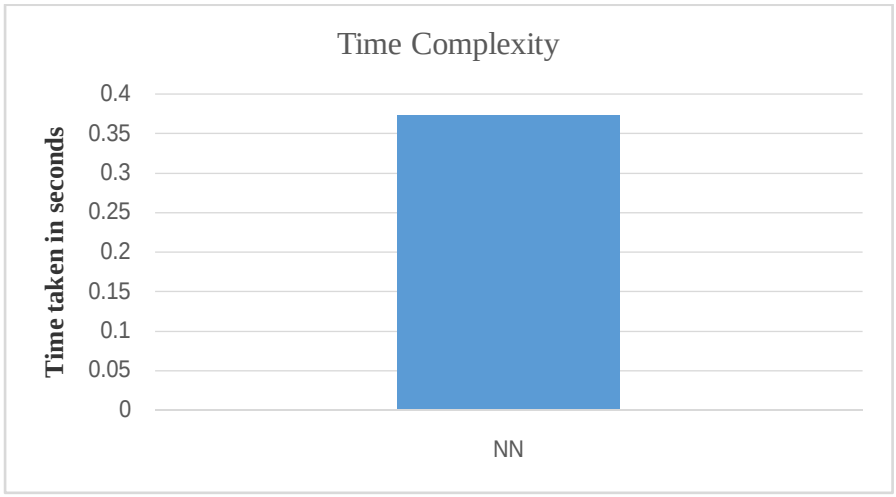


Figure 7: NN of Time Complexity

Time complexity is the difference between the time classified and submitted. In this portion, the time requirement of each and every individual image have been calculated and they are averaged as shown in fig.7. By calculating under this method, the overall time complexity is 0.3734%

Conclusion

This paper deals with the computer aided breast cancer identification based on the two different methods such as K-means algorithm and neural network. The K-means algorithm shows the tumor detected area result obtained from 12 different sample images. And also it is carried out at four different levels. In neural network technique, the data classification is obtained through entropy, mean, standard deviation, variance, co-variance and tumor detected area in sq.mm. On the whole, the method mentioned two methods reduced the time complexity of about 0.373403 which is very low and concentrated false positive ratio of neural network as well as accuracy of result when compared with the other computer aided classification methods on breast cancer. In the future will be implemented to clusters in the mammogram images.

References

[1] M. Bazzocchi and F. Mazzarella, “CAD systems for mammography: a real opportunity? A review of the literature,” http://www.springerlink.com/ content/x3157r8u72196h45/fulltext.pdf/, 2006.

[2] A. Oliver, “A new approach to the classification of mammographic masses and normal breasttissue,” in Proceedings of the 18th International Conference on Pattern Recognition, 2006,vol. 4, pp. 707 – 710.

[3] J. Wei, H-P. Chan, B. Sahiner, C. Zhou, and L. M. Hadjiiski, “Computer-aided detectionof breast masses on mammograms: Dual system approach with two-view analysis,” Med.Phys., vol. 36, no. 10, pp. 4451 – 4460, 2009.

[4] Gou, S., Yang, H., Jiao, L., Zhuang, X.: Algorithm of Partition Based Network Boosting for Imbalanced Data Classi_cation. The International Joint Conference on Neural Networks (IJCNN).1-6. IEEE (2010)

[5] Heang-Ping Chan, Jun Wei, Tao Wu and Mark A. Helvie; "Computer-aided Detection System for Breast Masses on Digital Tomosynthesis Mammograms: Preliminary Experience", Radiology, Vol. 237, No. 3, December 2005

[6] Nicholas Petrick, Berkman Sahiner and Lubomir M. Hadjiiski; "Breast Cancer Detection: Evaluation of a Mass-Detection Algorithm for Computer- aided Diagnosis Experience in 263 Patients", Radiology, Vol. 224, No. 1, July 2002.

[7] Reyer Zwiggelaar, James E. Schumm and Christopher J. Taylor; "The Detection of Abnormal Masses in Mammograms", Wolfson Image Analysis Unit, University of Manchester, Manchester, UK.

[8] Jun Liu, Xiaomig Liu, Jianxun Chen and J Tang. Improved Local Binary Patterns for Classification of Masses using Mammography. IEEE proceedings 2011; 2692-2695.

[9] Muhammad Hussain, Summrina Kanwal Wajid, Ali Elzaart and Mohammed Berbar. A Comparison of SVM Kernel Functions Breast

cancer Detection, Eighth International Conference computer Graphics, Imaging and Visualization IEEE 2011;

[10] F atima Eddaoudi, Fakhita Regrgui, AbdelhakMahmoudi and Najib Lamouri. Mass Detection Using SVM Classifier Based on Texture Analysis. Applied Mathematical Science 2011; 5: 367-379

[11] Leonardo de Oliveira Martins, Geraldo Braz Junior, Aristofanes Correa Silva, Anselmo Cardoso de Paiva and Marcelo Gattass. Detection of Masses in Digital Mammograms Using K-Means and Support Vector Machine. Electronics letters on Computer Vision and Image Analysis 2009; 8(2); 39- 50.

[12] B. N. Beena Ullala Mata and Dr. M. Meenakshi. A Novel Approach for Automatic detection of Abnormalities in Mammograms, IEEE proceedings 2011; 831-836.

[13] Nalini Singh, Ambarish G Mohapatra and Gurukalyan Kanungo. Breast Cancer Mass Detection in Mammograms Using K-means and Fuzzy C-means Clustering., International Journal of Computer Applications 2011; 22: 15-21.

[14] B. N. Beena Ullala Mata and Dr. M. Meenakshi. A Novel Approach for Automatic detection of Abnormalities in Mammograms, IEEE proceedings 2011; 831-836.

[15] Kemal Polat and Sahil Gunes. Breast Cancer Diagnosis using Least Square Support Vector machine. Digital Signal Processing 2007; 17: 694-701.

[16] Christos Stergiou and Dimitrios Siganos. Neural Network, Available on http://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol4/cs11/report.html