International Journal of Engineering & Technology, 7 (3.6) (2018) 154-159

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

A Comparative Study to Evaluate the Performance of Classification Algorithms in Mammogram Analysis

S.K. Sajan[1*], M. Germanus Alex[2]

[1]*Research Scholar, Research And Development Centre, Bharathiar University, Coimbatore, India.*
[2]*Assistant Professor And Head, Department Of Computer Science, Kamarajar Government Arts College, Surandai, Tirunelveli(Dt), Tamilnadu, India.E-Mail:Mgalxus@Yahoo.Com*
**Corresponding Author E-Mail:Sajan07041982@Gmail.Com*

Abstract

Breast cancer is a major threat humans are facing irrespective of geographical limits. The awareness about breast cancer has increased during the last decade and many preventive measures were in practice to detect the breast cancer before the symptoms were felt. Mammography is a screening methodology currently in practice. In this paper the mammogram image is analyzed using automated system. The automated system is designed to be capable of distinguishing the mammogram image into a normal or malignant. This process involves image enhancement and image segmentation at preprocessing level. Histogram equalization technique is used to transform low contrast region of the mammogram into region with higher contrast and Fuzzy C Means (FCM) algorithm is used to segment the mammogram image into regions suitable for further analysis. After enhancement and segmentation at preprocessing level the classification is done using three classification algorithms like decision tree classifier, Neural Network classifier and Support Vector Machine (SVM). The performance of the classification algorithms is evaluated using the following criteria like speed, flexibility, robustness, scalability, interpretability, Time complexity and also based on accuracy, sensitivity and specificity. The results obtained in classification are compared with other classification algorithms. It is found that the neural network classifier approach produces better results compared to other classifiers.The average accuracy in diagnosis by Neural Network approach classifier is around 91%. Also it is found that the decision tree approach is much flexible and easy to use compared to other approaches.

Keywords: *Image enhancement, mammogram, automated systems , benign, malignant, histogram equalization, robustness.*

1. Introduction

Breast cancer is a harmful disease that targets the females all over the world irrespective of their age. Mammogram[5] is a method to examine the breast for abnormalities by applying a low dose X-ray. It takes about 20 minutes time for completing a mammogram test. The results of the mammogram is copied directly into a computer for analysis, hence the digital images shows the cancer clearly compared with the X-ray film images. Digital mammography[17] is commonly in practice to detect breast cancer. There are two types of mammograms:

(i) Screening Mammogram
(ii) Diagnostic Mammogram

The screening mammogram is done for the women before the symptoms are felt. Diagnostic mammogram is done for women after the symptoms or a sign of breast cancer has been found. In this paper we perform analysis on only the screening mammogram[3].

Machine learning techniques are becoming the important tool for the Doctors/Radiologists to detect breast cancer at its early stage [3]. Predicting the breast cancer accurately requires an effective classification algorithm. Image classification is categorized into supervised and unsupervised, few supervised classification algorithms are Minimum distance, Bayes, Decision trees, Neural Network Classifier, Support Vector Machine (SVM) and Contextual Classifier. Also there are unsupervised classification algorithms like K-mean, Clustering, Fuzzy C Means and so on.

Classification [18] is the separation or ordering of objects into classes. A classification process in which classes have been pre-defined needs a method that will train the classification system to allocate objects to the classes. The training is based on a training sample, a set of sample data where for each sample the class is already known. We assume each object to have a number of attributes, one of which tells us which class the object belongs to. This attribute is known for the training data but for data other than the training data we assume that the value of the attribute is unknown and is to be determined by the classification method. When a new object is presented to the trained system it is able to assign the object to one of the existing classes this approach is called supervised learning. In supervised learning schemes, it is assumed that we have sufficient training data to build an accurate model during the training phase. The model that we build from the training data is never 100% accurate and classification based on the model will always lead to errors in some cases. In spite of such errors, Classification can be useful for prediction and better understanding of the data.

Aim and Objective

The main objective of this paper is to analyze the input image after mammogram using image processing techniques like image enhancement, segmentation and classification.

i) To avoid unnecessary biopsy due to false interpretation
ii) To identify the malignant tissues of the breast before the symptoms were felt.

Copyright © 2018 Authors. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

iii) Enhance the mammogram image to make the image suitable for further analysis.
iv) Segmentation of the mammogram images into different regions for analysis.
v) Classification of the mammogram image after segmentation using classification algorithms like decision tree classifier, SVM classifier and neural network classifier.

Limitations

i) Exposure to radiation frequently affects the health [2].
ii) Some cancers can't be found by a mammogram [5].

Signs and Symptoms of Breast Cancer

- A lump breast
- An unusual color of the milk especially bloody content
- Skin dimpling
- Dimpling or the distortion of the nipple
- Sore nipples or scaring

2. Literature Review

Support Vector Machine (SVM) for Classification

Support vector machine (SVM) is a supervised learning model used in the field of machine learning [7]. SVM builds one or more than one hyperplane in the high-dimensional feature space for classification. A hyperplane is applied to differentiate between data classes. If a hyperplane has the longest distance to the closest training data point of any class, then a beneficial division can be attained; since a larger margin denotes the classifier has a lower generalization error. When classes overlap with other, a hyperplane is chosen in order to minimize the errors of data points along or across the boundary line between classes; these points are referred to as the support points or support vectors. The support vectors thus specify the discrimination function clearly. The ability to identify the optimal discriminating hypersurface between two classes in a situation when many such separating hypersurfaces exist is a major strength of this approach and assists in dealing with the problem of overfitting during training[8].

Algorithm: SVM learning and classification[8]

1) Training: Select an appropriate kernel function, $k(\chi_i, \chi_j)$.
Interestingly, the choice of the kernel function is frequently not critical and classifier performance is often comparable for several of the simple kernel functions.
2) Minimize ||w|| subject to the constraint given in the equation $\omega_i(w.\chi_i - b) \geq 1 - \xi_i$. This isaccomplished by maximizing the kernel- modified Lagrange function, equation
$$L(w,b,a) = \sum_{i=1}^{N} a_i - \frac{1}{2}\sum_{i,j=1}^{N} a_i a_j \omega_i \omega_j k(\chi_i, \chi_j)$$
With respect to a_i subject to the constraints $\sum_{i=1}^{N} a_i \omega_i = 0$ and $a_i \geq 0$.
3) Store only the non-zero a_i and corresponding training vectors χ_i. These are the support vectors.
4) Classification : For each pattern x, compute the discrimination function, equation
$$f(x) = \sum_{i \in SV} a_i \omega_i k(\chi_i, x) + b.$$
using the support vectors χ_i and the corresponding weights a_i.
The sign of the function determines the classification of x.
The support vector machine divides the data classes into two linearly separable classes. In classification of mammogram image one class will be data associated with malignant and the other will be benign or normal.
In support vector machine the training can be based on a small set of support vectors[7]. Even in large training sets, thus limiting the computational complexity of training with explicitly represented feature vectors. The support vector classification algorithm is formally identical to that used for linear discrimination, while every dot product used earlier is replaced by a non-linear kernel function.

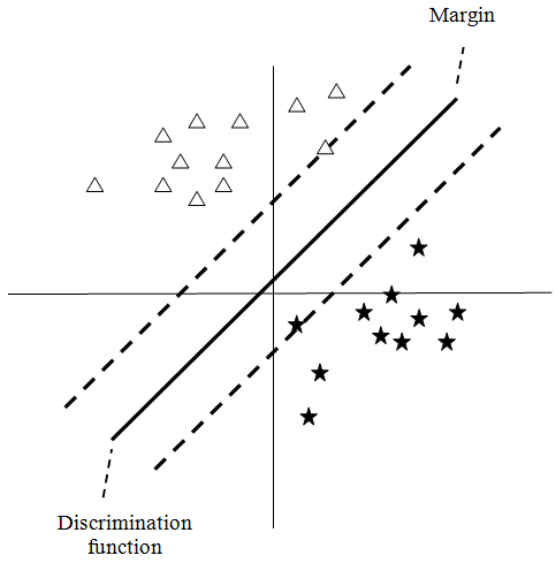


Fig. 1: Basic two-class classification idea of support vector machines

In non-linear separable approach to support vector machine classification, the input space having the classes as in Fig 2, that are not linearly separable can be converted to linearly separable by mapping the input space patterns to a new feature space in which the two classes are linearly separable as shown in the Fig 1.

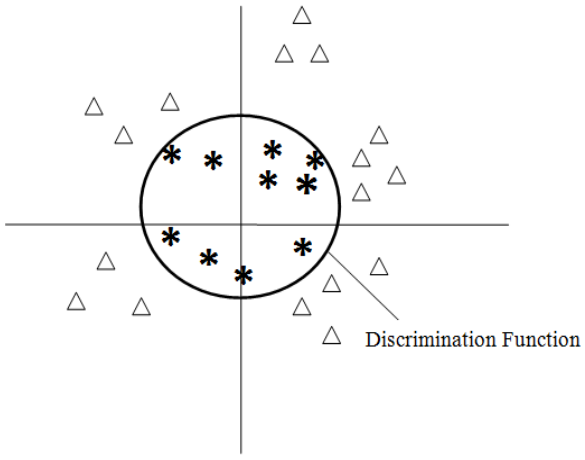


Fig. 2: Input space showing classes that are not linearly separable

Classification is a task of assigning an item to a certain category, which is called a class, based on the characteristic features ofthat item. This task in any classification system is performed by a classifier that takes a feature vector as an input and responds with a category to which the object belongs. A feature vector is a set of features extracted from the input data.

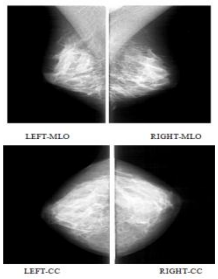


Fig. 3: Normal breast tissue

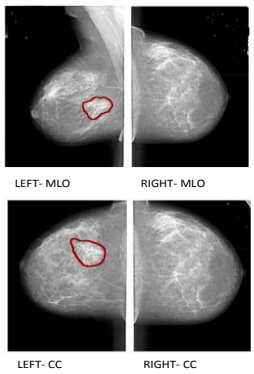


Fig. 4: Malignant breast tissue

Fig 3and Fig 4 shows the mammogram image showing normal, and malignant breast tissue respectively, from MIAS dataset[12] used especially for research purpose.

Neural Network Classifier

A Neural Network model [19] which is the branch of artificial intelligence is generally referred to as Artificial Neural Networks (ANNs). ANN teaches the system to execute task, instead of programming computational system to do definite tasks. To perform such tasks, Artificial Intelligence System (AI) is generated. It is a pragmatic model which can quickly and precisely find the patterns buried in data that replicate useful knowledge. One case of these AI models is neural networks. AI systems should discover from data on a constant basis. In the areas of medical diagnosis relationships with dissimilar data, the most available techniques are the Artificial Intelligence techniques. An artificial neural network is made up of many artificial neurons which are correlated together in accordance with explicit network architecture. The objective of the neural network is to convert the inputs into significant outputs.

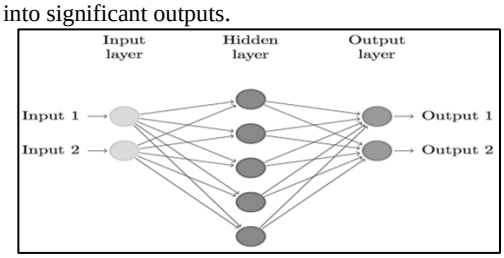


Fig. 5: Neural network classifier in mammogram image analysis

Decision Tree Classifier

This decision tree classifier is designed to classify the mammogram images which are efficient when dealing with large datasets and it is much easy to understand. The decision tree[18] approach of classification shows a tree structure similar to a flow chart. We need training data for classification in this method as shown in table 2.1.

Table 2.1: Classification Training Data

Image Name	Sore/Scaring Nipples	Lump breast	Skin dimpling	Dimpling/distortion of the nipple	Age>40	Class
Img001	Yes	Yes	Yes	No	No	Malignant
Img002	No	No	No	No	Yes	Benign
Img003	No	No	Yes	Yes	Yes	Malignant
Img004	Yes	No	No	No	No	Benign
Img005	No	Yes	No	No	Yes	Benign
Img006	No	No	No	No	No	Benign
Img007	No	No	No	No	No	Benign
Img008	Yes	No	No	Yes	Yes	Malignant
Img009	Yes	No	No	No	Yes	Benign
Img010	No	No	No	Yes	No	Benign
Img011	No	No	No	No	No	Benign
Img012	Yes	Yes	Yes	Yes	Yes	Malignant
Img013	No	No	Yes	No	No	Benign
Img014	Yes	No	No	No	Yes	Benign
Img015	No	No	Yes	No	Yes	Benign
Img016	No	No	Yes	No	No	Benign
Img017	Yes	No	No	No	Yes	Benign
Img018	No	No	No	No	Yes	Benign
Img019	No	No	No	No	Yes	Benign
Img020	No	Yes	No	No	Yes	Benign
Img021	No	No	No	No	Yes	Benign
Img022	Yes	No	No	No	No	Benign
Img023	No	No	No	No	Yes	Benign
Img024	No	Yes	No	No	Yes	Benign
Img025	No	No	No	No	No	Benign
Img026	No	No	Yes	No	Yes	Benign
Img027	No	Yes	No	No	No	Benign
Img028	Yes	Yes	Yes	No	Yes	Malignant
Img029	Yes	No	No	No	Yes	Benign
Img030	No	Yes	No	No	Yes	Benign

Each node of the tree in Fig 6 represents a decision while the leaves are the classes. It is both predictive and descriptive. The relationships found in the training dataset are projected by the decision tree. The training process that generates the tree is called induction. Usually the complexity of the decision tree increases as the number of attributes increases [18]. The quality of the decision tree depends on the quality of the training data. The quality of the decision tree is measured based on its accuracy in classification.

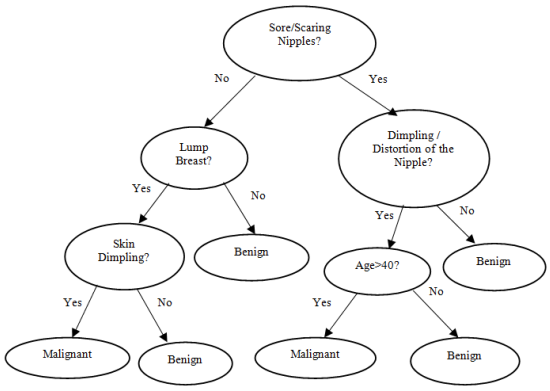


Fig. 6: Decision tree for the data in Table 2.1

Decision trees can lead to large errors if the number of training examples per class is small. This can happen when the tree has a number of classes and many attributes and the depth of the tree grows, So that towards the bottom of the tree the number of training samples is small leading to over-fitting the training data.

3. Method and Materials

The X-ray images of mammogram used in this study is from MIAS dataset [12]. Image enhancement and segmentation is performed at the pre-processing level to modify the X-ray image suitable for further analysis. Histogram equalization technique is used to transform low contrast region of the mammogram into region with higher contrast, if contrast is good the object will be visible. The degree of contrast of a mammogram is based on the features of the object and the imaging system. Image segmentation is done in the X-ray image of the mammogram using fuzzy C-Means segmentation algorithm to partition the X-ray image into parts or regions. This division into parts is often based on the characteristics of the pixels in the X-ray image of mammogram. For example, one way to find regions in an X-ray image is to look for abrupt discontinuities in pixel values, which typically indicate edges. Finally the classification algorithms like decision tree classifier, Neural Network classifier and SVM classifier is used to group the breast tissues into any one of the two categories benign or malignant.

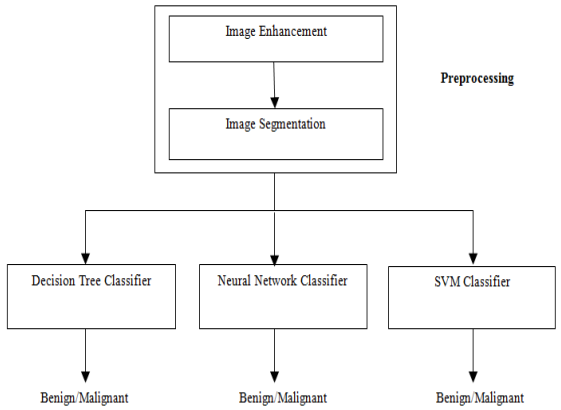


Fig. 7: Proposed method

Evaluate the performance of classification methods

The accuracy of a classification method is the ability of the method to correctly determine the class of a randomly selected data instance. The test dataset has a total of 25 mammogram images in each dataset represented as T objects. C is the correctly classified mammogram images from the T objects. The error rate is defined as $ErrorRate = \frac{(T-C)}{T}$. This error rate is only an estimate of the true error rate and it is expected to be a good estimate if the number of test data T is large and representative of the population. We now define accuracy, sensitivity and specificity as shown below:

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad (1.1)$$

$$Sensitivity = \frac{(TP)}{(TP+FN)} \quad (1.2)$$

$$Specificit\,y = \frac{(TN)}{(TN+FP)} \quad (1.3)$$

Where TP (True Positive) represents the number of cases who actually has breast cancer and was diagnosed correctly as malignant or abnormal, TN True Negative) represents the number of cases who actually has no breast cancer but was diagnosed incorrectly as malignant or abnormal, FP (False Positive) represents the number of cases who actually has no breast cancer but was diagnosed incorrectly as malignant or abnormal. FN (False Negative) represents the number of cases who actually has breast cancer and was diagnosed incorrectly as benign or normal. There are various other methods like bootstrapping, bagging and boosting that are used to improve the accuracy of classification results.
Some other evaluation criteria like speed, flexibility, robustness, scalability, interpretability and time complexity can be applied to get better understanding and predictive accuracy of the classification methods in mammogram analysis. Speed includes the time required to learn to use the model. Flexibility of a classification method is the ability to change based on the requirements. Robustness refers to the classification method being able to produce accurate results in spite of some errors and missing values in datasets. Scalability is required for the classification methods to work efficiently for large datasets. Interpretability in mammogram analysis is the ability to understand the results produced by the classification methods. Time complexity is the average time it takes to produce the results.

4. Experimental Results

The mammogram images from MIAS data set is taken for experimentation purpose and the result is used to evaluate the performance of the proposed method. Exactly four datasets with 25 mammograms in each data set are tested with the proposed approach and the result is given in table 4.1 and table 4.2.

Table 4.1: Results obtained by the Classification Methods

Classification Method	Accuracy for datasets					Sensitivity for datasets					Specificity for datasets				
	1	2	3	4	%	1	2	3	4	%	1	2	3	4	%
Decision Tree Classifier	81	87	90	88	87	75	80	78	70	76	90	70	74	71	76
Neural Network Classifier	80	94	91	97	91	80	80	83	86	82	90	95	94	95	94
SVM Classifier	77	85	83	91	84	70	76	78	75	75	88	98	90	90	92

The Average accuracy, sensitivity and specificity of decision tree classifier are 87%, 76% and 76% respectively. The performance of Neural Network classifier is good compared to other methods the accuracy, sensitivity and specificity of neural network classifier is 91%, 82% and 94% respectively. Also the support

vector machine produces the following results 84% accuracy, 75% sensitivity and 92% specificity.

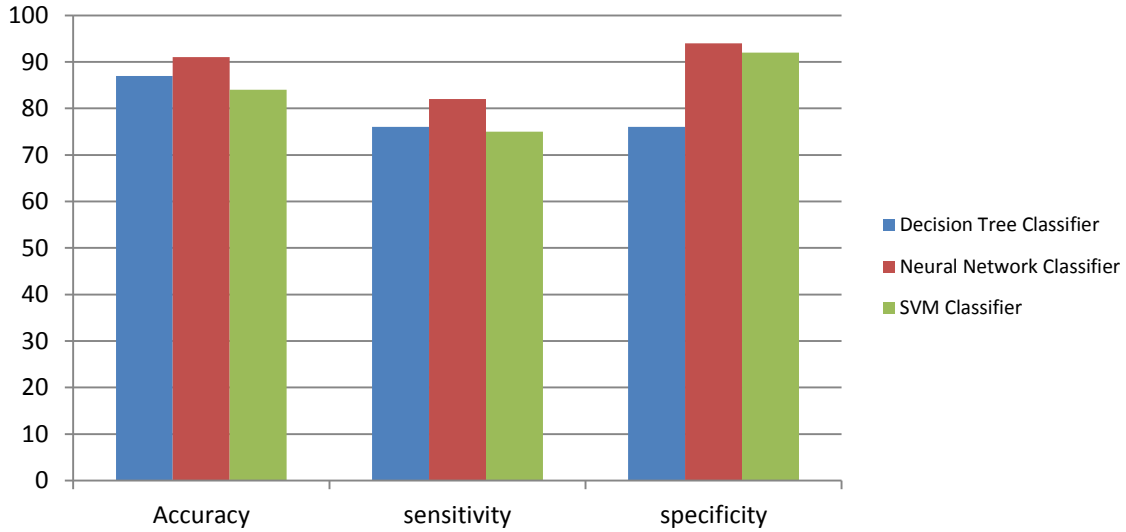


Fig. 8: Average percentage of accuracy, sensitivity and specificity represented pictorially

The pictorial representation of the table 4.1 is shown in the fig 8. The performance of neural network classifier is better and the specificity of SVM classifier is good and close to the value of the neural network classifier. At the same time the overall performance of SVM and decision tree classifiers are almost same with respect to accuracy and sensitivity.

Table 4.2: Results Obtained by Evaluating Certain Other Criteria

S.No	CRITERIA TO EVALUATE CLASSIFIERS	DECISION TREE CLASSIFIER	NEURAL NETWORK CLASSIFIER	SVM CLASSIFIER
1	Speed	0.98	0.98	0.9
2	Flexibility	0.99	0.85	0.88
3	Robustness	0.79	0.82	0.85
4	Scalability	0.82	0.82	0.81
5	Interpretability	0.7	0.97	0.91
6	Time Complexity	0.95	0.8	0.78
TOTAL		5.23	5.24	5.13

According to the table 4.1 and table 4.2 the overall performance of the Neural Network classifier is better compared with other classification algorithms. But the decision tree classifier is much flexible and easy to use. Time complexity is less in support vector machine classifier. The error rate is less in neural network classifier compared to other approaches. The speed and scalability of decision tree and neural network classifiers are the same.

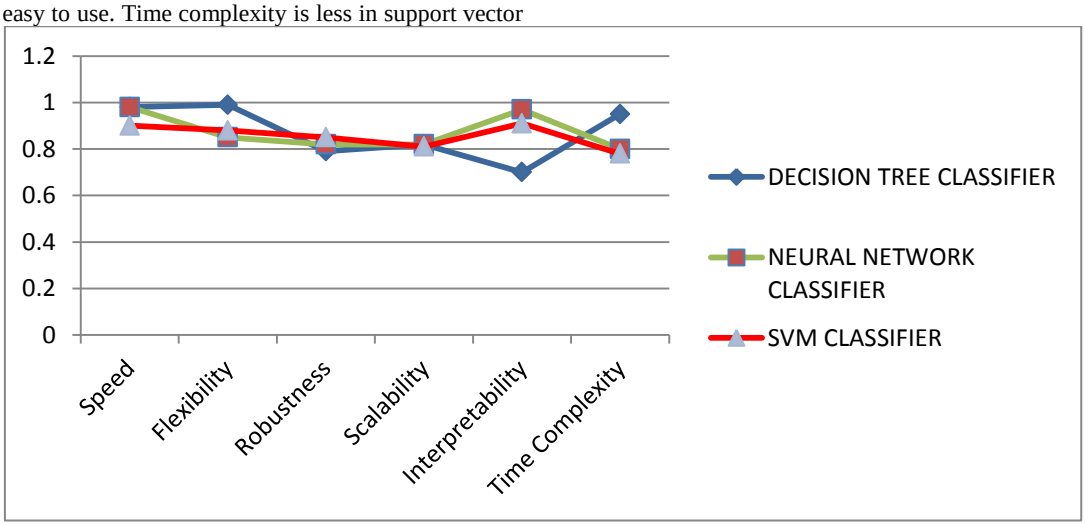


Fig. 9: Graphical representation of the table 4.2

5. Discussion and Conclusion

In this paper the accuracy of classification algorithms are analyzed based on various parameters. The problems in mammogram image analysis have been reduced by implementing automated systems to assist the radiologists, which reduce the unwanted biopsies and treatments. The performance of a classification algorithm in one dataset may vary with another dataset, hence selecting the appropriate classifier for the dataset is important. The predictive accuracy of three most efficient classification methods like decision tree, neural networks and SVM are evaluated in this study related with mammogram analysis. The results obtained in

classification shows that the neural network classifier produces good results compared to other classification methods and the decision tree method is much flexible, fast and easy to use than all other classification methods. The neural network classifier can be used to test its performance in other areas like remote sensing, advanced medical imaging, microscopic imaging and so on.

References

[1] www.cancer.org/breast-cancer-early-detection
[2] www.cancer.org/cancer/breastcancer/.../breastcancerearlydetection/breast
[3] www.cancer.gov/cancertopics
[4] Gonzalez RC, Woods RE & Eddins SL, *Digital Image Processing Using MATLAB*,2nd Edition, McGraw Hill Education (India) Private Limited, (2013).
[5] http://www.breastcancerindia.net/bc/statistics
[6] Gonzalez RC & Woods RE, *Digital Image Processing*, Third Edition, Pearson Education, Inc, (2009).
[7] Cortes C & Vapnik V, "Support-vector networks", *Machine learning*, Vol.20, No.3, (1995), pp.273–297.
[8] Sonka M, Hlavac V & Boyle R, "Digital Image Processing and Computer Vision", *Cengage Learning*, (2011).
[9] Chanda B & Dutta Majumder D, *Digital Image Processing and Analysis*, Prentice Hall of India Private Limited, (2007).
[10] Shinghal R, *Pattern Recognition Techniques and Applications*, Oxford University Press, (2008).
[11] Joshi MA, *Digital Image Processing An Algorithmic Approach*, PHI Learning Private Limited, (2009).
[12] www.mias.org
[13] www.birads.at
[14] Eddaoudi F, Regragui F, Mahmoudi A & Lamouri N, "Masses Detection Using SVM Classifier Based on Textures Analysis", *Applied Mathematical Sciences*, Vol.5, No.8, (2011), pp.367-379.
[15] Ema T, Doi K, Nishikawa RM, Jiang Y & Papaioannou J, "Image feature analysis and computer-aided diagnosis in mammography: reduction of false-positive clustered microcalcifications using local edge-gradient analysis", *Med. Phys.*, Vol.22, No.2, (1995), pp.161–169.
[16] Enderwich CY & Tzanakou EM, "Classification of Mammographic tissue using shape and texture features", *Proceedings of the 19th International Conference-IEEE/EMBS*, (1997), pp.810–813.
[17] Eltoukhy MM, Faye I & Samir BB, "A comparison of wavelet and curvelet for breast cancer diagnosis in digital mammogram", *Computers in Biology and Medicine*, Vol.40, (2010), pp.384–391.
[18] Gupta GK, *Introduction to Data Mining with Case Studies*, Second Edition, PHI Private Ltd, (2011).
[19] Saravanan K & Sasithra S, "Review on classification based on artificial neural networks", *International Journal of Ambient Systems and Applications (IJASA)*, Vol.2, No.4,(2014), pp.11-18.