www.ijecs.in
International Journal Of Engineering And Computer Science ISSN: 2319-7242
Volume 4 Issue 9 Sep 2015, Page No. 14187-14193

Classification of Breast Mass classification - CAD System with Performance Evaluation

Dr. S. Mohan Kumar, Dr G. Balakrishnan

Associate Professor, Department of Computer Science and Engineering,
New Horizon College of Engineering, Bangalore,
Karnataka, India
E-Mail: mohankumar.sugumar@gmail.com

Director,
Indra Ganesan College of Engineering,
Trichy, Tamil Nadu, India

***Abstract* - Mammogram is measured the most consistent method for early detection of breast cancer. Computer-aided diagnosis system is also able to support radiologist to detect abnormalities earlier and more rapidly. In this paper the performance evaluation of the computer aided diagnostic system for the classification of mass classification in digital mammogram based on Discrete Wavelet Transform (DWT), Stochastic Neighbor Embedding (SNE) and the Support Vector Machine (SVM) is presented. This proposed system classifies the mammogram images into normal or abnormal, and the abnormal severity into benign or malignant. Mammography Image Analysis society (MIAS) database is used to evaluate the proposed system. The average classification rate achieved is satisfied.**

Keywords: DWT, SNE, MIAS, SVM,SSNE,Mammography.

INTRODUCTION

Breast cancer is the most extensive cancerous pathology among women. It is also an important public health problem in the world. As causes of its onset are still unknown, there are no efficient ways to avoid breast cancer. For this reason, an efficient diagnosis in its early stage can give women a better chance of full healing and survival. Therefore, early detection of breast cancer is the key for reducing the associated morbidity and death rates.

To study the human breast, Mammography is widely used as a diagnostic and a screening tool that uses X-rays. The objective of mammography is the premature revealing of breast cancer, usually through detection of characteristic microcalcifications and/or masses. Mammography is the only effective and viable technique to detect breast cancer in particular in the case of minimal tumors. About 30% to 50% of breast cancers reveal deposits of calcium called microcalcifications. Computer aided diagnosis system is also able to support radiologist to detect abnormalities earlier and faster.

1.1 Related Research Works:

All the following mentioned related research works are reviewed aptly to construct the proposed system with the high efficiency, A Computer Aided Diagnosis (CAD) system for the automatic detection of clustered microcalcifications in digitized mammograms is presented by (Song yang Yu, 2000). A computerized scheme for detecting early stage microcalcification clusters in mammograms is proposed by (Ryohei Nakayama, 2006). A computer aided decision support system for an automated diagnosis and classification of breast tumor using mammogram is presented by (M. Suganthi, 2009). A new method of feature extraction from Wavelet coefficients for classification of digital mammograms is proposed by (Ibrahima Faye, 2009). A novel methodology for the classification of suspicious areas in digital mammograms is presented by (Peter Mc Leod, 2010), and so on.

In this research the proposed system uses, two techniques for building a computer aided diagnostic system for the classification of microcalcification in digital mammograms based on DWT and SNE are presented. The SNE applied to wavelet transformed image and also applied on sub-bands of wavelet transformed image individually. SNE is essentially used for reducing high dimensionality data into relatively low dimensional data, efficiently. Then classifier system based on Support Vector Machine (SVM) is constructed. Experiments are conducted on Mammography Image Analysis society (MIAS) database. The MIAS is an organization of UK research groups interested in the understanding of mammograms and has generated a database of digital mammograms. Films taken from the UK National Breast Screening programme have been digitized to 50 micron pixel edge with a Joyce -Loebl

scanning microdensitometer. It is a device linear in the optical density range 0 to 3.2 and representing each pixel with an 8-bit word. MIAS database consists of a total of 322 digital mammogram images (161 breast pairs) in the mediolateral oblique view. The performance of the proposed system is carried on 99 normal images and 25 microcalcification images. Among the 25 abnormal images, there are 12 benign and 13 malignant images available. All the images are considered for the classification test.

METHODOLOGY

The proposed system for the classification of microcalcification in digital mammograms is built based on DWT, SNE and by applying SVM for building the classifiers, PCA for Comparison. In this following section the theoretical background of all the approaches are introduced.

A: Support Vector Machines (SVM)

SVMs are a set of related supervised learning methods that analyze data and recognize patterns, used for classification and regression analysis. The standard SVM is a non-probabilistic binary linear classifier, i.e. it predicts, for each given input, which of two possible classes the input is a member of. A classification task usually involves with training and testing data which consists of some data instances. Each instance in the training set contains one "target value" (class labels) and several "attributes" (features). SVM has an extra advantage of automatic model selection in the sense that both the optimal number and locations of the basic functions are automatically obtained during training. The performance of SVM largely depends on the kernel.

SVM is essentially a linear learning machine. For the input training sample set

$$(x_i, y_i), i = 1 \ldots . n, x \in R^n, y \in \{-1, +1\}$$

Let the classification hyperplane equation is to be $(\omega . x) + b = 0$ (1)

Thus the classification margin is 2 / |ω| .To maximize the margin, that is to minimize |ω|, the optimal hyperplane problem is transformed to quadratic programming problem as follows,

$$\begin{cases} min\emptyset(\omega) = \frac{1}{2}(\omega, \omega) \\ s.t. y_i\big((\omega . x) + b\big) \geq 1, \ \ i = 1,2 \ldots . l \end{cases} \quad (2)$$

After introduction of Lagrange multiplier, the dual problem is given by,

$$\begin{cases} max\ Q(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_i y_j \alpha_i \alpha_j K\big(x_i . x_j\big) \\ s.t \ \sum_{i=1}^{n} y_i \alpha_i = 0, \ \ \alpha_i \geq 0, i = 1,2 \ldots , n \end{cases} \quad (3)$$

According to Kuhn-Tucker rules, the optimal solution must satisfy

$$\big(y_i\big((w . x_i) + b\big) - 1 = 0, i = 1,2, .. n \quad (4)$$

That is to say if the option solution is

$$\alpha^* = (\alpha_1^*, \alpha_2^*, \ldots . , \alpha_i^*)^T, i = 1,2, \ldots n \quad (5)$$

Then

$$* w \ = \sum_{i=1}^{n} \alpha_i^* y_i x_i \quad (6)$$

$$b^* = y_i - \sum_{i=1}^{n} y_i \alpha_i^* \big(x_i . x_j\big) \ , j \in \{j | \alpha_i^* > 0\} \quad (7)$$

For every training sample point x_i , there is a corresponding Lagrange multiplier. And the sample points that are corresponding to $\alpha_i = 0$ don't contribute to solve the classification hyper plane while the other points that are corresponding to $\alpha_i > 0$ do, so it is called support vectors. Hence the optimal hyper plane equation is given by,

$$\sum_{x, \in SV} \alpha_i y_i \big(x_i . x_j\big) + b \ = 0 \quad (8)$$

The hard classifier is then,

$y = sgn\big[\sum_{x, \in SV} \alpha_i y_i \big(x_i . x_j\big) + b\big]$ situation, SVM constructs an optimal separating hyperplane in the high dimensional space by introducing kernel function $K(x . y) = \emptyset(x) . \emptyset(y)$, hence the nonlinear SVM is given by,

$$\begin{cases} min\emptyset(\omega) = \frac{1}{2}(\omega, \omega) \\ s.t. y_i \Big(\big(\omega . \emptyset(x_i)\big) + b\Big) \geq 1, i = 1,2, \ldots . l \end{cases} \quad (9)$$

And its dual problem is given by,

$$\begin{cases} maxL(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} y_i y_j \alpha_i \alpha_j K(x_i . x_j) \\ s.t. \sum_{i=1}^{n} y_i \alpha_i = 0, \ \ 0 \leq \alpha_i \leq C, i = 1,2, \ldots , l \end{cases} \quad (10)$$

Thus the optimal hyperplane equation is determined by the solution to the optimal problem.

B: Discrete Wavelet Transform (DWT)

Nowadays, wavelets have been used quite frequently in image processing and used for feature extraction, denoising, compression, face recognition, and image super-resolution. The decomposition of images into different frequency ranges permits the isolation of the frequency components introduced by "intrinsic deformations" or "extrinsic factors" into certain sub-bands. This process results in isolating small changes in an image mainly in high frequency sub-band images.

The 2-D wavelet decomposition of an image is performed by applying 1-D DWT along the rows of the image first, and, then, the results are decomposed along the columns. This operation results in four decomposed sub-band images referred to as low–low (LL), low–high (LH), high–low (HL), and high–high (HH).

C: Stochastic Neighbor Embedding (SNE)

SNE is a probabilistic approach to the task of placing objects, described by high-dimensional vectors or by pair-wise dissimilarities in a low-dimensional space in a way that preserves neighbor identities. A Gaussian is centered on each object in the high-dimensional space and the densities under this Gaussian (or the given dissimilarities) are used to define a probability distribution over all the potential neighbors of the object. The aim of the embedding is to approximate this distribution as well as possible when the same operation is performed on the low-dimensional "images" of the objects. A natural cost function is a sum of Kullback-Leibler divergences, one per object, which leads to a simple gradient for adjusting the positions of the low-dimensional images.

For each object, i and each potential neighbor, j the asymmetric probability is calculated by the formula that i would pick j as its neighbor is given by

$$pij = \frac{\exp(-d_{ij}^2)}{\sum_{k \neq i} exp(-d_{ik}^2)} \qquad (11)$$

The dissimilarities, d_{ij}^2 , may be given as part of the problem definition (and need not be symmetric), or they may be computed using the scaled squared Euclidean distance ("affinity") between two high-dimensional points, X_i; X_j :

$$d_{ij}^2 = \frac{\|X_i - X_j\|^2}{2\sigma_i^2} \qquad (12)$$

Where σ_i is either set by hand or found by a binary search for the value of σ_i that makes the entropy of the distribution over neighbors equal to $\log k$. Here, k is the effective number of local neighbors or "perplexity" and is chosen by hand. In the low-dimensional space, the Gaussian neighborhoods are used with a fixed variance so the *induced* probability qij that point i picks point j as its neighbor is a function of the low-dimensional *images* y_i of all the objects and is given by the expression

$$qij = \frac{exp\left(-\|y_i - y_j\|^2\right)}{\sum_{k \neq i} exp\left(-\|y_i - y_j\|^2\right)} \qquad (13)$$

The aim of the embedding is to match these two distributions as well as possible. This is achieved by minimizing a cost function which is a sum of Kullback-Leibler divergences between the original (pij) and induced (qij) distributions over neighbors for each object is given by (4)

$$C = \sum_i \sum_j pij \; log \frac{pij}{qij} = \sum_i KL(P_i||Q_i) \qquad (14)$$

The minimization of the cost function in Equation 4 is performed using gradient method. The gradient has the simple form as

$$\frac{\partial C}{\partial Y_i} = 2 \sum_j (y_i - y_j)\left(p_{ij} - q_{ij} + p_{ij} - q_{ij}\right) \qquad (15)$$

The gradient descent is initialized by sampling map points randomly from an isotropic Gaussian with small variance that is center around the origin. For speed up the optimization and avoid been stuck in local optima, a momentum term is added to the gradient [4]. The current gradient is added to an exponentially decay sum of previous gradients in order to determine the changes in the coordinates of the map points at each iteration of gradient search.

$$y^{(t)} = y^{(t-1)} \eta \frac{\partial J}{\partial y_i} + \alpha(t)\left(y^{(t-1)} - y^{(t-2)}\right) \qquad (16)$$

Where $y^{(t)}$ indicate the solution at iteration t, η indicates the learning rate, and $\alpha(t)$ represents the momentum at iteration t. In the early stages of the optimization, after the each iteration, a random jitter is added to the map points. Then gradually reducing the variance of this noise performs a type of simulated annealing that helps the optimization to escape local minima in the cost function.

D: Principle Component Analysis (PCA)

Given a set of data on n dimensions, PCA aims to find a linear subspace of dimension d lower than n such that the data points lie mainly on this linear subspace .Such a reduced subspace attempts to maintain most of the variability of the data. The linear subspace can be specified by d orthogonal vectors that form a new coordinate system, called the `principal components'. The principal components are orthogonal, linear transformations of the original data points, so there can be no more than n of them.

However, the hope is that only $d < n$ principal components are needed to approximate the space spanned by the n original axes. The most common definition of PCA is that, for a given set of data vectors x_i $i \in 1,...t$, the d principal axes are those ortho normal axes onto which the variance retained under projection is maximal. In order to capture as much of the variability as possible, let us choose the first principal component, denoted by U_1, to have maximum variance. Suppose that all centered observations are stacked into the columns of a $n \times t$ matrix X , where each column corresponds to an n -dimensional observation and there are t observations. Let the first principal component be a linear combination of *X* defined by coefficients (or weights) $w = w_1 ..... w_2$.

In matrix form:

$$U_1 = w^T X$$

$$\text{var}(U_1) = \text{var}(w^T X) = w^T S w$$

where S is the $n \times n$ sample covariance matrix of X . Clearly $\text{var}(U_1)$ can be made arbitrarily large by increasing the magnitude of w Therefore, w is chosen in order to maximize $w^T S w$ while constraining w to have unit length.

$$\max w^T S w$$

$$subject\ to\ w^T w = 1$$

DOI: 10.18535/ijecs/v4i9.20

To solve this optimization problem a Lagrange multiplier α_1 is introduced:

$$L(w,\alpha)=w^T Sw-\alpha_1(w^T w-1) \quad (17)$$

Differentiating with respect to w gives n equations,

$$Sw=\alpha_1 w$$

Pre-multiplying both sides by w^T, we get

$$w^T Sw=\alpha_1 w^T w=\alpha_1$$

$\mathrm{var}(U_1)$ is maximized if α_1 is the largest Eigen value of S .Clearly α_1 and *w* are an Eigen value and an eigenvector of S. Differentiating (5.5) with respect to the Lagrange multiplier α_1 gives us back the constraint:

$$w^T w=1$$

This shows that the first principal component is given by the normalized eigenvector with the largest associated Eigen value of the sample covariance matrix S. A similar argument can show that the d dominant eigenvectors of covariance matrix S determine the first *d* principal components. Another nice property of PCA, closely related to the original discussion is that the projection onto the principal subspace minimizes the squared reconstruction error,

$$\sum_{1=1}^{t}\left\|x_i-\hat{x}_i\right\|^2 . \quad (18)$$

In other words, the principal components of a set of data in $\Re^n$ provide asequence of best linear approximations to that data, for all ranks $d\le n.$

Consider the rank-d linear approximation model as:

$$f(y)=\bar{x}+U_d y \quad (19)$$

This is the parametric representation of a hyper plane of rank d .

For convenience, suppose $\bar{x}=0$ (otherwise the observations can be simply replaced by their centered versions $\tilde{x}=x_i-\bar{x}.$ Under this assumption the rank d linear model would be $f(y)=U_d y$ where U_d is a $n\times d$ matrix with d orthogonal unit vectors as columns and y is a vector of parameters. Fitting this model to the data by least squares leaves us to minimize the reconstruction error:

$$\min_{U_d,y_i}\sum_{i}^{t}\left\|x_i-U_d y_i\right\|^2 \quad (20)$$

By partial optimization for y_i we obtain:

$$\frac{d}{dy_i}=0\Rightarrow y_i=U_d^T x_i \quad (21)$$

Now we need to find the orthogonal matrix U_d :

$$\min_{U_d}\sum_{i}^{t}\left\|x_i-U_d U_d^T x_i\right\|^2 \quad (22)$$

Define $H_d=U_d U_d^T$. H_d is a $n\times n$ matrix which acts as a projection matrix and projects each data point x_i onto its rank d reconstruction. In other words, $H_d x_i$ is the orthogonal projection of x_i onto the subspace spanned by the columns of U_d A unique solution U can obtained by finding the singular value decomposition of $X.$ For each rank d , U_d consists of the first d columns of U .Clearly the solution for U can be expressed as singular value decomposition (SVD) of X (J. Friedman,2002).

$$X=U\Sigma V^T \quad (23)$$

Since the columns of U in the SVD contain the eigenvectors of XX^T . Figure 1 and 2 shows the histogram plot for normal mammogram and benign and malignant mammograms images using PCA based dimension reduction. The histogram plots show the variation in the benign and malignant pattern.

III PCA Vs SNE

To analyze the performance of the SNE, the proposed system is tested with the state of art technique Principal Component Analysis (PCA) using SVM classifier. Initially the dimension of the feature is reduced for the wavelet transformed image and the performance is analyzed. In the second approach, dimension reduction is applied on the wavelet sub-bands individually and the classification accuracy is calculated. The classification accuracy obtained by the proposed system using SVM classifier for first stage and final stage is shown in Table 1 and 2. Figure 1 and 2 show the histogram plot for normal mammogram and benign and malignant mammograms images using SSNE based dimension reduction.

DOI: 10.18535/ijecs/v4i9.20

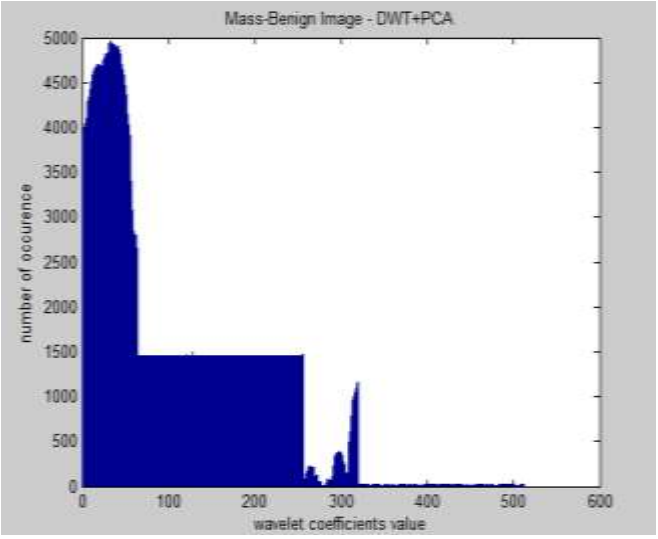
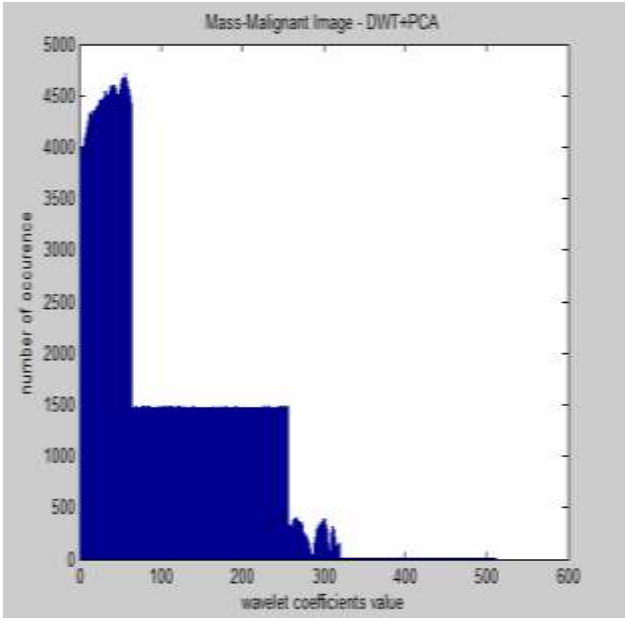


Figure 1 Histogram plot of dimension reduced 2-level wavelet coefficients of a mass- benign image by PCA



The maximum average classification accuracy of 93.39% is achieved by SNE at initial while using the wavelet decomposed image. It is observed that the

Table 2 Classification results of proposed mass classification system for final stage based on PCA and SNE using SVM classifier

Mass - Benign/Malignant classification				
Level of decomposition	Wavelet		Wavelet Sub-bands	
	PCA	SNE	PCA	SNE

The SNE classified the abnormal images more than 10% than PCA reduction technique. The classification accuracy of the final stage classifier based on wavelet sub-band features, the PCA reduction techniques produces less than 90% average accuracy while SNE produces a

Figure 2 Histogram plot of dimension reduced 2-level wavelet coefficients of a mass- malignant image by PCA

Table 1 Classification results of proposed mass classification system of first stage based on PCA and SNE using SVM classifier

Normal/Abnormal classification				
Level of decomposition	Wavelet		Wavelet Sub-bands	
	PCA (%)	SNE (%)	PCA (%)	SNE (%)
2	88.31	90.84	86.37	85.25
3	88.70	91.22	82.45	89.05
4	90.10	89.94	84.59	89.44
5	85.40	88.55	84.47	88.28
6	88.31	**93.39**	84.59	86.88

performance of SNE is somewhat better than PCA based reduced features. Table 5.8 shows the performance of PCA and SNE for benign/malignant classification.

	(%)	(%)	(%)	(%)
2	87.98	88.12	82.86	92.10
3	78.81	92.10	84.07	92.10
4	85.42	90.75	84.21	90.75
5	86.70	92.10	84.21	**93.39**
6	84.07	90.83	84.14	89.47

maximum average of 93.39% as shown in bold values. Hence it is concluded from the tables that the proposed SNE based method outperforms the PCA method in all aspect which is very clear in Figures 3 and 4.

DOI: 10.18535/ijecs/v4i9.20

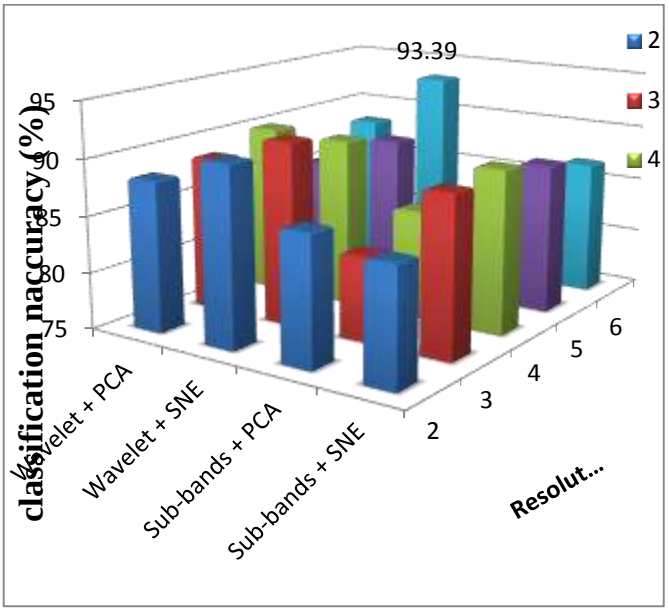


Figure 3 Graphical representations of performance of normal/abnormal mass classification using PCA and SNE

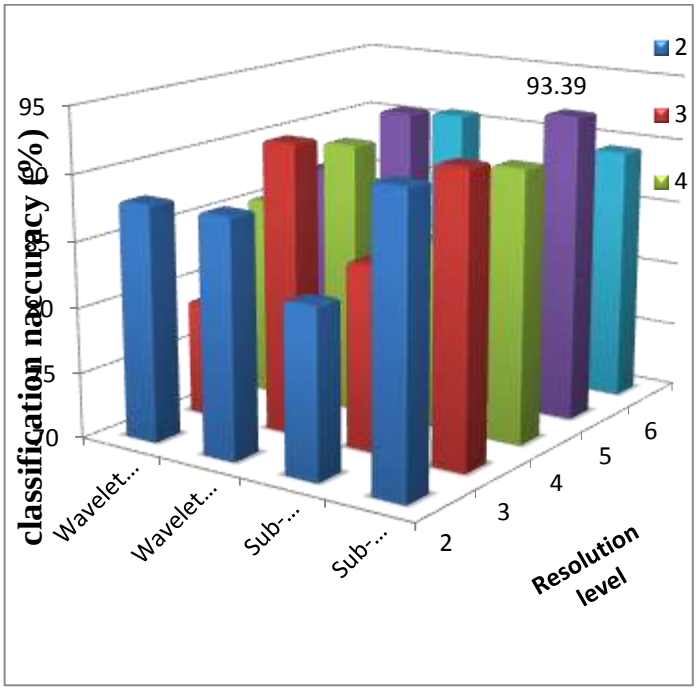


Figure 4 Graphical representations of performance of benign/malignant mass classification using PCA and SNE

The classification accuracy shows in the above table gives the accuracy of SNE and PCA for mass classifications severity into benign or malignant. The abnormal severity is correctly classified with no error by the SNE reduced data set for wavelet at all the level decomposition and wavelet sub-bands techniques at 5^{th} level decomposition. The maximum accuracy obtained by SNE is at all the level decomposition and applied on the wavelet sub-bands individually. The bar chart shows in the Figure 3 and 4 clearly shows the performance of SNE over PCA. The most of the highest bars in the charts are belongs to SNE that shows the efficiency of SNE over PCA.

CONCLUSION:

This proposed system classifies the mammogram images into normal or abnormal, and the abnormal severity into benign or malignant. The proposed methods are implemented in MATLAB and the performances of these methods are also analyzed productively. Finally, in order to serve the cancer patients with due diagnosis, the classification accuracy rate is sensibly derived from our proposed system.

REFERENCES

[1] Songyang Yu and Ling Guan, "A CAD System for the Automatic Detection of Clustered Microcalcifications in Digitized Mammogram Films", IEEE Transactions on Medical imaging, vol. 19, no. 2, February 2000, pp 115-126.

[2] Ryohei Nakayama and Yoshikazu Uchiyama, "Computer-Aided Diagnosis Scheme Using a Filter Bank for Detection of Microcalcification Clusters in Mammograms", IEEE Transactions on Biomedical engineering, vol. 53, no. 2, February 2006, pp 273-283.

[3] M.Suganthi and M.Madheswaran, "Mammogram Tumor Classification using Multimodal Features and Genetic Algorithm", IEEE International Conference on "Control, Automation, Communication and Energy conservation, June 2009, pp 1-6.

[4] Ibrahima Faye and Brahim Belhaouari Samir, "Digital Mammograms Classification Using a Wavelet Based Feature Extraction Method", IEEE conference on Computer and Electrical Engineering, 2009, pp 318-322.

[5] Peter Mc Leod and Brijesh Verma, "A Classifier with Clustered Sub Classes for the Classification of Suspicious Areas in Digital Mammograms", IEEE conference on Neural Networks, July 2010, pp 1-8.

[6] Viet Dzung Nguyen, Thu Van Nguyen and Tien Dzung Nguyen, "Detect Abnormalities in Mammograms by Local Contrast Thresholding and Rule-based Classification", IEEE third International Conference on Communications and Electronics, August 2010, pp 207-210.

[7] Andy Tirtajaya and Diaz D. Santika, "Classification of Microcalcification Using Dual-Tree Complex Wavelet Transform and Support Vector Machine", IEEE International Conference on Advances in Computing, Control and Telecommunication Technologies, December 2010, pp 164-166.

[8] Fatemeh Saki and Amir Tahmasbi, "A Novel Opposition-based Classifier for Mass Diagnosis in Mammography Images", IEEE Iranian Conference of Biomedical Engineering, November 2010, pp 1-4.

[9] Alireza Shirazi Noodeh and Hossein Rabbani, "Detection of Cancerous Zones in Mammograms using Fractal Modeling and Classification by Probabilistic Neural Network" IEEE Iranian Conference of Biomedical Engineering, November 2010, pp 1-4..

[10] K. Thangavel and A. Kaja Mohideen, "Semi-Supervised K-Means Clustering for Outlier Detection in Mammogram Classification", IEEE Trendz in Information Sciences & Computing, December 2010, pp 68-72.

[11] Mohamed Meselhy Eltoukhy and Ibrahima Faye, "Curvelet Based Feature Extraction Method for Breast Cancer Diagnosis in Digital Mammogram", IEEE International Conference on Intelligent and Advanced Systems, June 2010, pp 1-5.

[12] Dheeba.J and Tamil Selvi.S, "Classification of Malignant and Benign Microcalcification Using SVM Classifier", IEEE International Conference on Emerging Trends in Electrical and Computer Technology, March 2011, pp 686-690.

[13] Y.Ireaneus Anna Rejani, S.Thamarai Selvi "Early detection of breast cancer using SVM classifier technique" International Journal on Computer Science and Engineering, Vol 1(3), 2009, 127-130.

[14] Smola A. J., Scholkopf B., and Muller K. R., "The connection between regularization operators and support vector kernels", Neural Networks New York, vol.11, November 1998, pg 637-649.

[15] MIAS:database, http://peipa.essex.ac.uk/ipa/pix/mias/.